Development of computational tools for modeling the biotransport of small organic molecules into the active site of broad-substrate specificity enzymes

A Dissertation
Submitted to the Faculty of the
University of Minnesota
By

Diego Ernesto Escalante Pérez

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

Alptekin Aksan, Advisor
Lawrence P. Wackett, Co-Advisor

July, 2019

— To all who helped make this a a reality —

Abstract

In this dissertation research, two new computational tools were developed to model the biotransport of small organic molecules into the active site of broad-substrate specificity (BSS) enzymes. The biological organism selected to develop, test and validate these tools were Rieske non-heme iron dioxygenases. Members of this family of enzymes are known to have biocatalytic activity on more than three hundred different substrates. The large diversity of substrates that can be acted upon makes these enzymes very attractive in biotechnological processes such as bioremediation. In addition, the highly specific chirality of the products obtained makes these enzymes attractive for the potential synthesis of pharmaceutical precursors. Currently, the most common way to identify new substrates requires formulating an educated guess followed by the arduous task of testing each possible compound individually. This slows down the pace at which new industrial processes can be formulated or current ones further developed. The tools presented in this research provide fundamental and practical scientific contributions.

For the basic science studies of my dissertation, an all-atom and, a coarse-grained (CG) model of Rieske non-heme iron dioxygenases were used to investigate the factors that affect the biotransport of small organic molecules into their active sites. From the all-atom model I discovered a gating mechanism that allows aromatic substrates into the active site and blocks other compounds. The key to these gates are T-stacked $\pi - \pi$ interactions between hydrophobic amino acids and the aromatic substrates. On the other hand, from the CG model I discovered that the shape of tunnel modulates the hydrophobicity level of the surface. As the tunnels become more concave, the hydrophobicity increases causing the formation of a water exclusion zone which increases the diffusivity of aromatic substrates. The CG models also revealed that convex tunnels prevent the adhesion of hydrophobic substrates to the tunnel walls; providing a possible explanation for the evolution of bottlenecks at the entrance of Rieske active sites.

For the practical contributions of my dissertation, I developed two new computational tools for the prediction of Rieske substrates. The first tool is an all-atom algorithm that models the stochastic roto-translational movement of small organic molecules along the Rieske enzyme tunnels. This algorithm has a 92% prediction accuracy of Rieske substrates. In addition, it is capable of elucidating the location of high-energy barriers along the tunnel, allowing the formulation of possible protein engineering sites. The second tool is a CG non-dimensional model of the Rieske enzyme tunnels. This algorithm has a 90% prediction accuracy of Rieske substrates. The processing time of 1ms/substrate combined with its high accuracy allows for the high-throughput screening of possible Rieske substrates.

List of Publications

I) Aukema, K. G., **Escalante, D. E.**, Maltby, M. M., Bera, A. K., Aksan, A., & Wackett, L. P. (2016). In silico identification of bioremediation potential: carbamazepine and other recalcitrant personal care products. *Environmental science & technology*, 51(2), 880-888. doi:10.1021/acs.est.6b04345

II) **Escalante, D. E.**, Aukema, K. G., Wackett, L. P., & Aksan, A. (2017). Simulation of the bottleneck controlling access into a Rieske active site: predicting substrates of naphthalene 1, 2-dioxygenase. *Journal of chemical information and modeling*, 57(3), 550-561. doi:10.1021/acs.jcim.6b00469

III) **Escalante, D. E.**, & Aksan, A. (2019). Role of water hydrogen bonding on transport of small molecules inside hydrophobic channels. *The Journal of Physical Chemistry B.* doi:10.1021/acs.jpcb.9b03060

IV) **Escalante, D. E.**, & Aksan, A. (2019). Prediction of ligand transport along hydrophobic enzyme Nanochannels. *Computational and Structural Biotechnology Journal.* 17, 757-760. doi:10.1016/j.csbj.2019.06.001

Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction

1.1 Motivation

Enzymes are capable of catalyzing almost every known chemical reaction.[1] Enzymes are now used worldwide in several industries such as food, agriculture, chemical synthesis, renewable energy, and medicine.[2] Enzyme mediated processes are rapidly becoming the standard because of their lower energy input, cost effectiveness, non-toxicity, and eco-friendliness.[3,4] It is projected that the global enzymes market will be valued at $10.4 billion by the year 2024.[5]

However, new enzymatically-catalyzed reactions must be discovered for this projection to be met.[5] To achieve this goal, the first task is to determine the right enzyme/substrate pair; nonetheless, this remains an overwhelming task.[6] Experimentally, thousands of hours would need to be devoted to this task, increasing the risk of exposure to toxic, carcinogenic or mutagenic compounds.[7,1,8,9,10,11,12] Computationally, the currently available tools require prohibitively expensive simulations that hinder the possibility of high-throughput screening.[13,14,15,16,17,18]

In my dissertation research, I developed, tested, and validated new computational tools that can rapidly identify substrates of enzymes belonging to the Rieske non-heme iron dioxygenase family. The models were based on the premise that for biocatalysis to occur, the substrates (ligands) must be transported along access tunnels into the active sites of the Rieske enzymes.[19,20,21,22,23,24] Therefore the models take into consideration: i) the structural dynamics of the enzyme; ii) the mechanism

of ligand entry into the active site; iii) the non-bonded interactions between the enzyme and the ligand as it moves along the access tunnel; and iv) the configurational space energetics of the ligand once it has reached the active site. Since members of the Rieske family are broad-substrate specificity (BSS) enzymes[25] the models developed have the potential of identifying thousands[7] of different substrates that can be biotransformed into: i) non-toxic products[26,27,28,29,30] (bioremediation); ii) products with higher commercial value than the substrate[31,6] (biocatalysis) – as illustrated in Figure 1.1. Therefore, the computational methods presented in this dissertation will be a valuable tool for the rational prediction of novel substrates for the production of biofuels, food and agricultural additives, and pharmaceuticals precursors.

1.2 Background Theory

1.2.1 Enzyme catalysis

Many chemical reactions occur spontaneously and instantaneously; others require catalysts to increase the rate of the chemical reaction. The catalyst provides a reaction pathway with a lower activation energy barrier to convert the substrate into a product. Thermodynamically, this is equivalent to reducing the free energy of activation, while keeping constant the initial and final free energy states (Figure 1.2). In addition, catalysts are not consumed during the reaction. In principle, this means that they could be used unlimited times; in practice this usually never happens. The lifetime of

catalysts is limited by its ability to avoid being inhibited, deactivated or destroyed.[32]

Enzymes are biocatalysts that have evolved over millions of years to efficiently perform all the biochemical reactions needed to sustain life. Even more astounding is the possibility that there is a corresponding enzyme that can catalyze any given chemical reaction under mild biological conditions.[1] Nonetheless, finding the correct enzyme/ligand combination becomes a gargantuan task, both experimentally and computationally.

Catalysis in enzymes with buried active sites is a complex process that involves multiple steps. In this dissertation, I solely focus on the first two steps: i) transport of the substrate into the core of the enzyme; and ii) proper positioning of the substrate inside the active site. The transport of substrates has been shown to occur through the use of access tunnels connecting the buried active site to the surface of the enzyme.[33,34,18,35] These tunnels provide the enzyme with mechanisms that: i) allow preferential access to substrates; ii) prevent the generation of toxic intermediates that can damage of cellular organelles;[36] iii) enable reactions that require dry environments;[37] and iv) synchronize reactions involving multiple substrates.[38,39] The second step can proceed only if transport of the substrate through the tunnel is successful.

It has been hypothesized that certain enzymes families owe their broad-substrate specificity (BSS) to the access tunnels.[24] When the active sites are large vestibules that can accommodate a wide range of ligand sizes,[20,40,41] the tunnels serve as the primary mechanism to limit the transport of ligands. This is advantageous for organ-

isms because it reduces the number of enzymes needed to perform the same type of chemical reaction in substrates with similar properties.

1.2.2 Force fields and nonbonded interactions

In molecular modeling, a force field is the mathematical expression used to calculate the total potential energy of a system. The expression is a parameterized analytical way to describe the intra- and inter- molecular interaction potential energy $U(\mathbf{r}_1, \mathbf{r}_2 \ldots \mathbf{r}_N)$. There are various ways of calculating the parameters of a force field including the use of ab initio quantum mechanical calculations[42] or fitting them to experimental data from neutron, X-ray and electron diffraction.[43] Each molecule is defined simply as a set of spheres (atoms) that are held together by simple elastic (harmonic) forces. There are many force fields available in the literature having different degrees of complexity and each aim to treat different kinds of systems. Some of the most commonly used force fields include AMBER, CHARMm, and OPLS. The typical expression for a force field is given by Equation 1.1:

$$\begin{aligned} U = & \sum_{\text{bonds}} \frac{1}{2} k_b (r - r_0)^2 + \sum_{\text{angles}} \frac{1}{2} k_a (\theta - \theta_0)^2 \\ & + \sum_{\text{torsions}} \frac{V_n}{2} \left[1 + \cos(n\phi - \delta)\right] + \sum_{\text{improper}} V_{imp} \\ & + \sum_{\text{Coulomb}} \frac{q_i q_j}{r_{ij}} + \sum_{\text{LJ}} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \end{aligned} \tag{1.1}$$

The first four terms quantify the *intra*molecular or local contributions to the total energy, and the last two the *inter*molecular nonbonded interactions. During the transport process of a ligand through the enzyme channel, covalent bonds are not formed. Therefore, **this dissertation focused on the effect that nonbonded forces have on the transport properties of ligands into the active site of enzymes**. This reduces Equation 1.1 to:

$$U = \sum_{\text{Coulomb}} \frac{q_i q_j}{r_{ij}} + \sum_{\text{LJ}} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{1.2}$$

Force fields divide the nonbonded interactions into two: i) electrostatic interactions; and ii) van der Waals interactions. In Equation 1.2, the first type of nonbonded interaction serves to describe the electrostatic (Coulomb) interactions between atoms. This type of interaction quantifies the force between two stationary and electrically charged particles. Although a molecule may be formally neutral, the asymmetric distribution of electrons causes a partial charge distribution, i.e., a dipole moment. The partial charges can be derived by a fitting experimental thermodynamic data.[44,43,45] However, *ab initio* calculations provide the most reliable partial charges.[43]. The second type of nonbonded force arises from the attractive and repulsive van der Waals interaction between atoms. The repulsive force is due to the overlap in the electron clouds of both atoms and generally scales at a short-range described by the r^{12} term. On the other hand, the interactions between induced dipoles result in an attractive component that scales at a long-range and is described by the r^6 term.

One of the most common ways to quantify the van der Waals interactions is the 12-6 Lennard Jones potential. This potential defines a set of atomic parameters, ε and σ, that are used to quantitatively describe these types of nonbonded interactions. The term ε is the well-depth that measures how strongly the two particles attract each other. And the term σ is the distance at which the intermolecular potential between the two particles is zero. This term is also referred to as the van der Waals radius as it determines how close the two nonbonded particles can get before they start repelling each other. The stability of an arrangement of atoms is a function of Lennard-Jones separation distance r. Figure 1.5 shows that as the distance between atoms decrease below the equilibrium distance the potential energy rapidly increases indicating a repulsive force due to the overlap of atomic orbitals. On the other hand, at a distance greater than the atoms start to experience an attractive force. For the calculation of the pairwise potential energy between two atoms, the Berthelot combining rules (Equations 1.3 and 1.4) are used:

$$\varepsilon_{ij} = \sqrt{\varepsilon_i \times \varepsilon_j} \tag{1.3}$$

$$\sigma_{ij} = \frac{\sigma_i + \sigma_j}{2} \tag{1.4}$$

The nonbonded forces, shown in Equation 1.2, govern the transport properties of ligands.[43].

1.2.3 Molecular Dynamic Simulations

1.2.3.1 Principle

An Molecular Dynamics (MD) simulation is a technique used to produce a dynamical trajectory for a system composed of N particles by integrating Newton's equations of motion. The setup of a MD simulation requires: i) a set of initial conditions (positions and velocities); ii) a force field definition; iii) an ensemble definition. Then the classical equation of motion, given in Equation 1.5, is solved:

$$m_i \frac{\mathrm{d}^2 \mathbf{r_i}}{\mathrm{d}t^2} = \mathbf{f_i} = -\frac{\partial}{\partial \mathbf{r}_i} U\left(\mathbf{r}_1, \mathbf{r}_2 \ldots \mathbf{r}_N\right) \tag{1.5}$$

where $U\left(\mathbf{r}_1, \mathbf{r}_2 \ldots \mathbf{r}_N\right)$ is the potential energy depending on the coordinates of N particles. This coupled system of N second order non-linear differential equations must be solved numerically.

1.2.3.2 Initial Conditions

The initial position and velocities of all N particles in the system must be specified. In the case of MD simulations involving real enzymes and ligands, the coordinates of all atoms can be obtained from X-ray crystal structures. The coordinates are available from the Protein Databank (PDB). In the case of model systems, the three-dimensional coordinates must be generated to represent the desired structure. The

initial velocity of each particle is generated from a Maxwell-Boltzmann distribution, given by Equation 1.6:

$$p_{\mathrm{MB}}(v) = 4\pi v^2 \left(\frac{m}{2\pi k_{\mathrm{B}} T}\right)^{3/2} e^{\frac{-mv^2}{2k_{\mathrm{B}}T}} \tag{1.6}$$

where m is the mass of the particle, v is the velocity, k_{B} is the Boltzmann constant and T is the absolute temperature. The velocities obtained from the probability distribution are adjusted to result in a net zero angular momentum about the center of mass of the system.

1.2.3.3 Evaluation of Forces

Calculation of individual pairwise interaction forces is simply the partial derivate of the potential with respect to the three-dimensional Cartesian coordinates, $\mathbf{r}$. The calculation of the interaction forces due to the Lennard-Jones potential is demonstrated in Equation:

$$\begin{aligned} \mathbf{f_{ij,vdW}} &= -\frac{\partial}{\partial \mathbf{r}_{ij}} \left\{ 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} \right] \right\} \\ &= \frac{48\varepsilon}{\sigma_{ij}^2} \left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{14} - \frac{1}{2}\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{8} \right] \mathbf{r}_{ij} \end{aligned} \tag{1.7}$$

where ε_{ij} and σ_{ij} are obtained using Equations 1.3 and 1.4, respectively.

1.2.3.4 Integration algorithms

The equation of motion (1.5) must be solved numerically. Therefore, the trajectory must be discretized and an integrator must be used over small time steps to simulate the passage of time:

$$\mathbf{r}_i(t_0) \rightarrow \mathbf{r}_i(t_0 + \Delta t) \rightarrow \mathbf{r}_i(t_0 + \Delta 2t) \rightarrow \ldots \rightarrow \mathbf{r}_i(t_0 + \Delta nt) \tag{1.8}$$

The integrator function used to advance time must have the following properties:

- Minimal need to compute forces. This is the most time consuming step.
- Good stability if large Δt are used.
- High accuracy.
- Conservation of energy and momentum.
- Conservation of phase space volume (i.e., simpletic integrators)

The four most common time integrators are:

i) Taylor expansion(Equation 1.10);

ii) Verlet (Equations 1.11-1.12);

iii) leap-frog (Equations 1.13-1.14); and

iv) velocity-Verlet (Equations 1.15-1.16).

$$\mathbf{a}_i(t_0) = \frac{\mathrm{d}^2\mathbf{r}_i(t_0)}{\mathrm{d}t^2} \tag{1.9}$$

$$\mathbf{r}_i(t_0 + \Delta t) = \mathbf{r}_i(t_0) + \frac{\mathrm{d}\mathbf{r}_i(t_0)}{\mathrm{d}t}\Delta t + \frac{1}{2}\frac{\mathrm{d}^2\mathbf{r}_i(t_0)}{\mathrm{d}t^2}\Delta t^2 + O(\Delta t^3) \tag{1.10}$$

$$\mathbf{r}_i(t_0 + \Delta t) = -\mathbf{r}_i(t_0 - \Delta t) + 2\mathbf{r}_i(t_0) + \mathbf{a}_i(t_0)\Delta t^2 + O(\Delta t^3) \tag{1.11}$$

$$\mathbf{v}_i(t_0) = \frac{1}{2\Delta t}\left[\mathbf{r}_i(t_0 + \Delta t) - \mathbf{r}_i(t_0 - \Delta t)\right] \tag{1.12}$$

$$\mathbf{r}_i(t_0 + \Delta t) = \mathbf{r}_i(t_0) + \mathbf{v}_i\left(t_0 + \frac{\Delta t}{2}\right)\Delta t \tag{1.13}$$

$$\mathbf{v}_i\left(t_0 + \frac{\Delta t}{2}\right) = \mathbf{v}_i\left(t_0 - \frac{\Delta t}{2}\right) + \mathbf{a}_i(t_0)\Delta t \tag{1.14}$$

$$\mathbf{r}_i(t_0 + \Delta t) = \mathbf{r}_i(t_0) + \mathbf{v}_i(t_0)\Delta t + \frac{1}{2}\mathbf{a}_i(t_0)\Delta t^2 \tag{1.15}$$

$$\mathbf{v}_i(t_0 + Deltat) = \mathbf{v}_i(t_0) + \frac{1}{2}\left[\mathbf{a}_i(t_0) + \mathbf{a}_i(t_0 + \Delta t)\right]\Delta t \tag{1.16}$$

1.2.4 Rieske non-heme Iron Dioxygenases

Rieske non-heme iron dioxygenases are a family of broad-substrate specificity (BSS) enzymes known to biotransform thousands of different anthropogenic xenobiotics and recalcitrant contaminants.[46,21,47,27,29,25,7] The crystal structures of all available Rieske enzymes reveal the presence of multiple tunnels.[25,7,48] The longest of these tunnels connects the enzymes' buried active sites to the cytoplasm-exposed surface.[20,21,22] The tunnels' geometries, hydrophobicities, polarities, charges, and dynamics are expected to affect the transport properties of ligands into the active sites of Rieske enzymes.[23,49]

Another important structural feature of Rieske enzymes is the presence of the catalytic mononuclear iron (i.e., the non-heme center) at the distal end of the tunnel. The mononuclear iron is anchored to the enzyme by two histidine residues and a carboxylate-containing residue, also known as the 2-His-1-carboxylate facial triad[50] (Figure 1.3). At the beginning of the catalytic cycle, an electron transfer chain is established via the facial triad.[42,51] The electron chain yields a highly reactive iron species with a low activation energy barrier.[42,51] The mononuclear iron is now ready to donate one of its electrons in order to complete the catalytic cycle.[28] Since the highly reactive iron species can donate an electron to any aromatic compound present inside the active site,[52] a full catalytic event can be interpreted as a successful transport of the ligand through the tunnel.

1.3 Background Information

1.3.1 Predicting Enzymatic Catalysis

Different models have been proposed over the last century in order to explain enzymatic catalysis, from the early *lock–key model* to the *induced-fit* model and, as recently as 2012, the *keyhole–lock–key model*.[53] On the basis of a recent report that over 60% of annotated enzymes have their active site buried within the protein core,[24] the keyhole–lock–key model appears to be the one that can best represent enzymatic catalysis. In addition, for this model it has been proposed that recognition of a substrate by the enzyme is a two-step process: i) access of the ligand into the buried active site via a connecting channel,[53,54] and ii) complementary fit of the ligand to the active site.[53]. The use of a dual process leads to a greater possibility of correct substrate identification by the enzyme. This is probably an evolutionary feature to prevent access of nonpreferred or inhibitory chemicals,[55,56] prevent damage to the cell through the production of highly reactive intermediates,[36] make reactions that require the absence of water possible or quasi-vacuum states kinetically and thermodynamically feasible,[37] and ensure correct timing of reactions requiring a precise order of steps.[39,57] All of these benefits can protect the cell, but it makes the computational identification of substrates and characterization of products a nontrivial task, as two distinct and computationally intensive processes have to be analyzed.

Most substrate (and inhibitor) prediction methods focus on studying the second part of the discrimination process and implicitly consider any complicated steps in the conversion of the substrate by fitting the model parameters to experimental kinetic data.[58] The quantitative structure—activity relationship (QSAR) was one of the earliest methods to quantitatively model the enzymatic catalysis process and predict substrates.[59,60] In this method, physiochemical properties of the ligand are correlated to the experimentally determined bioactivity of a given enzyme.

As the available computational power increased, molecular docking was developed, allowing studies of the complementarity between the three-dimensional structures of the target enzyme and the ligand.[61] This technique uses the coordinates of the ligand and the receptor (enzyme active site) to find the most energetically favorable conformation and position of the ligand.[61] Several scoring functions have been developed to estimate the binding affinity of the ligand upon finding the best conformation, and many of the available docking protocols allow for induced-fitting effects by varying rotable bonds of both the ligand and receptor.[61,62] It should be emphasized that the methods described above focus on active-site binding and do not consider the access tunnel space.

Access tunnels, also refered to as channels, are dynamic structures that undergo changes in shape, size, and orientation within a short time scale,[53] upon natural breathing,[33,54] or in response to water or ligand passage,[54] and these changes can be studied via atomic-scale fluctuations or domain-scale motions.[53] Classical Molecular Dynamics simulations are commonly used to explore and identify these motions, but

by themselves they cannot identify access channels, and are rarely used to model transport of ligands into/out of the active site. To identify access channels, several computational tools such as POCKET,[63] LIGSITE,[64] CAVER,[65] and MOLE 2.0.[23] have been developed. Another "cost-effective" approach that has been used for the identification of channels leading into the active site is the sliding box docking approach.[66,67] Although the sliding box method uses short MD simulations ($\sim$ 40ns), it is still limited by the fact that only low-energy ligand binding poses are identified, and it cannot identify potential energy barriers.[67] For enzymes in which the physiological time scale at which channels open (approximately microseconds to milliseconds)[68] is longer than the available computational time scale of classical MD simulations ($\sim 10^2$ns), techniques such as random acceleration molecular dynamics (RAMD)[16] and steered molecular dynamics (SMD)[14,16,34,35,54,69] can be used.

1.3.2 Modeling Ligand Transport along Enzyme Channels

1.3.2.1 Equilibrium Methods

Using equilibrium MD (eMD) simulations, the transport of ligands through a tunnel can be described as a discrete sequence of ligand binding/unbinding events where each step has an equilibrium probability $P(x)$:

$$P(x) = \frac{e^{-\beta U(x)}}{\int e^{-\beta U(x)}\, dx} \tag{1.17}$$

where $\beta = k_{\mathrm{B}}T$. Figure 1.4 shows conceptually various enzyme/ligand dissociation processes starting from a single bound state and finishing in different unbound states. The probability, $P(x)$, is very low for each binding/unbinding event along the tunnel (shown as black dots). Nonetheless, transitions through all of these regions of low probability are theoretically possible in ergodic systems. Therefore, if equilibrium MD simulations were run for an infinite amount of time every state would be accessed allowing for the calculation of all ligand transport properties.

Using eMD simulations is highly impractical due to the large size of the system (10^5 atoms). A large system require long computational times to perform all the necessary force calculations.[18] Unfortunately, such a large system is required because water molecules must be included to properly capture the dynamics of both the ligand and the enzyme.[18] In addition, the tunnels may dynamically open and close in response to water passage. The tunnel dynamics can also fluctuate in response to the breathing motion of the enzyme itself, caused by the inevitable thermal fluctuations in the system.[70] Due to these factors, there are no reports of eMD simulations being used to model the transport of a ligand from the active site to the surface, or vice versa. Therefore, nonequilibrium trajectories must be computed in order to stay within the computational timescale constraints.

1.3.2.2 Nonequilibrium Methods

The two most common nonequilibrium methods to simulate ligand transport are: steered molecular dynamic (SMD) simulations and, accelerated molecular dynamics (aMD). For these methods it is convenient to introduce the term "reaction coordinate" (RC) to describe the transitions between bound and unbound states. The RCs consist of the sampled nonequilibrium intermediate structures that connect the bound and unbound states.

Steered molecular dynamics (SMD) is a special type of MD simulation that applies an external force vector to a ligand in order to study its egress path from the active site to the bulk environment. The expulsion force used in SMD has the following form:

$$\mathbf{f} = -\frac{1}{2}k\nabla\left[vt - (\mathbf{r} - \mathbf{r}_0)\mathbf{n}\right]^2 \tag{1.18}$$

where k is a spring constant, v is the constant pulling speed of the ligand, r and r_0 are the current and initial positions of the pulled ligand, respectively, and $\mathbf{n}$ is the unit vector specifying the pulling direction of the ligand. One of the drawbacks of SMD is that the pulling direction is kept constant throughout the entire simulation. A second limitation is that the force needs to be identified *a priori*. This means that if the pulling force is too strong, i.e., to reduce the computational time, the movement of the ligand might induce unnatural distortions to the structure of the enzyme resulting in incorrect estimations of the diffusion and thermodynamic parameters.[18]

The first application of SMD simulations to study the movement of ligands through channels was performed by the Schulten and co-workers back in 2003.[71] In this study they found that ligand transport along the tunnels of the glycerol kinases found in *Escherichia coli* is controled by the formation and breaking of hydrogen bonds.[71] More recently, Skovstrup et. al.[72] used SMD to elucidate the substrate transport pathway of the GABA receptors in the human body. And Fukunishi et. al[73] have shown that the egress of ligands involved in the hydroxylation of Vitamin D3 in Cytorchome P450s is controled by salt briges that form along the access tunnels. There are many other studies that have used SMD to elucidate key feactures affecting the transport properties of small organic molecules along enzyme channels.[74,75,76,77,78,79] However, in all of these studies caution must be excised because the applied external force might have caused un-natural rearrangements of the enzyme channels geometry.

In contrast to SMD, in aMD simulations the ligand is pulled out of the enzymes' active sites by applying randomly oriented small magnitude forces, $\mathbf{n}$, at different points along the tunnel. The multiple forces applied cause the ligand to wander around the binding cavity, allowing it to chose its own direction to exit the enzyme. In brief, the direction $\mathbf{n}$ is chosen so that $\mathbf{f} = f\mathbf{n}$ where f is the constant magnitude of the randomly chosen force. This force is maintained for a predefined number of simulation steps m during which the velocity of the ligand is calculated and compared against a threshold value given by $v = \tau/mdt$, where dt is the time step and τ is the minimum distance before the direction is changed. If the velocity of the ligand at a given steps does not overcome the threshold it indicates that the direction of movement must be changed as a steric collision point has been reached. This allows

the calculations to simulate infrequent unbinding events without prior knowledge of the conformation states, potential energy wells, or barriers that the ligand might encounter along the enzyme tunnel. Some of the advantages of this type of method are that: i) it speeds up the timescale by orders of magnitude; ii) it allows for finding alternative ligand diffusion routes; and iii) no initial prediction of the exit pathway is needed.[18] However, the main disadvantage still remains. A full nanosecond scale simulation must be performed for each ligand, limiting the scalability of this method.

1.3.3 Nonbonded Interactions Inside the Active Site

1.3.3.1 Molecular Docking

One of the most widely used methods to understand the thermodynamic interactions between a small molecule, such as a ligand, and an enzyme is molecular docking[80–82]. The goals of the docking programs is to semiquantitatively rank the binding ability of small molecules to a specified conformation of the enzyme active site pocket[80]. As it was shown in Figure 1.6 small molecules are able to adopt several different conformations with different energy levels. The docking algorithms generally sample all these poses inside of the rigid active site receptor by sampling the nonbonded interactions. Several docking studies have been successfully performed in the past in order to generate databases of lead compounds for the development of new drugs[83].

1.3.3.2 Predicting Substrates Based on Active Site Interactions

Similar to NDO, the cytochrome P450 (CYP450) enzyme family has been extensively studied experimentally because of its ability to metabolize xenobiotics.[58] In addition, since five isoforms of CYP450 can metabolize $\sim$ 90% of the marketed drug compounds,[84] there has been a lot of interest in computational modeling and prediction of substrates and inhibitors. The large body of literature on CYP450 prediction methods provides a benchmark to compare new prediction methods (albeit on a different enzyme class) such as the one presented in this dissertation. The CYP450 prediction methods can be broadly categorized as ligand-based methods and structurebased methods.[58]

Some of the most common ligand-based methods that have been used for the prediction of CYP450 substrates are linear partial least-squares,[85] neural networks,[86,87] recursive partitioning,[88] MACCS keys and FP4 fingerprints,[89] and support vector methods.[90,91] Similar to our prediction methodology for NDO, all of the previously mentioned methods showed a prediction accuracy of $> 88\%$. On the other hand, all of these ligand-based methods rely on statistical learning to develop a prediction model and thus require a training set with a sufficiently large sample size to develop a robust classification system.[90] The use of large training data sets can suffer from interlaboratory variations in experimental protocols that significantly affect the consistency of the data sets[90] and limit the model to a confined applicability domain based on the chemicals on which the training set was built.[58] Since our model did not require a training data set, it was not confined to a narrow applicability do-

main, which allowed us to do a less constrained exploration of the chemical space for investigating the very diverse substrates of NDO.

The second category of CYP450 prediction methods is the structure-based approach, in which the three-dimensional structure of the enzyme is used to determine how the ligand fits into the active-site cavity.[58] Thus, our prediction model method falls under this category. Hritz et al.[92] showed that with docking into the original apo crystal structure of $CYP450_{2D6}$, their prediction accuracy was $\sim 20\%$. Furthermore, they also showed that generating more conformations of the enzyme through MD simulations improved the prediction accuracy to 52%. A study by Teixeira et al.[93] showed that docking into 125 MD-generated structures yielded only 20% accuracy in their prediction of substrates of $CYP450_{3A4}$ substrates, Hayes et al.[94] had to use induced-fit docking to allow the protein to adapt to the ligand structure.

1.4 Dissertation Objectives and Overview

The main objective of my dissertation was to develop computational tools that could evaluate the thermodynamic properties and model the biotransport of small organic molecules into the active site of broad substrate specificity enzymes. The development of these tools allowed will be beneficial for the prediction of substrates of Rieske non-heme iron oxygenases, as well as other BSS enzyme families.

In order to achieve my main goal, the research was partitioned into four specific

aims:

S.A. 1 Develop an all-atom Monte Carlo method to model biotransport of ligands into active site of BSS enzymes.

S.A. 2 Identify enzymes for the biodegradation of emerging recalcitrant pollutants.

S.A. 3 Develop building blocks to model solvation effects in the biotransport of ligands along cylindrical and non-cylindrical nanochannels.

S.A. 4 Integrate building blocks for the development of a non-dimensional model to predict the biotransport of ligands along BSS enzyme nanochannels.

These specific aims were addressed in Chapters 2, 3, 4, and 5 of this dissertation, respectively. Chapter 6 discusses the need of understanding the nonbonded interactions that happen inside the active site of BSS enzymes. Chapter 7 outlines the RAPID website (rapid.umn.edu) for publishing the predictions made by the developed models. A summary of the research in provided in Chapter 8.

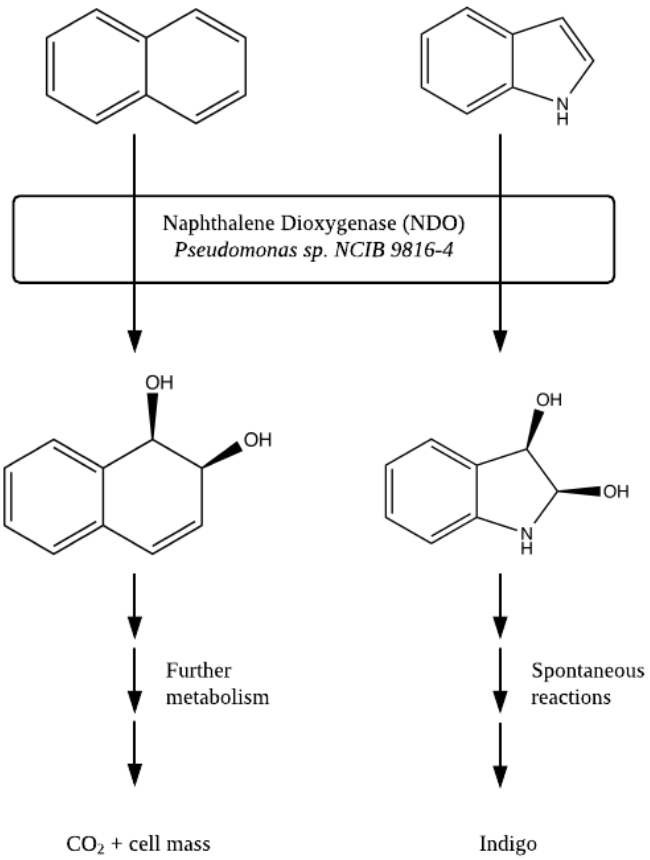

Figure 1.1: Biodegradative and biocatalytic reactions catalyzed by Naphthalene 1,2-dioxygenase from *Pseudomonas sp.* NCIB-9816. Figure adapted from Wackett et. al.[6]

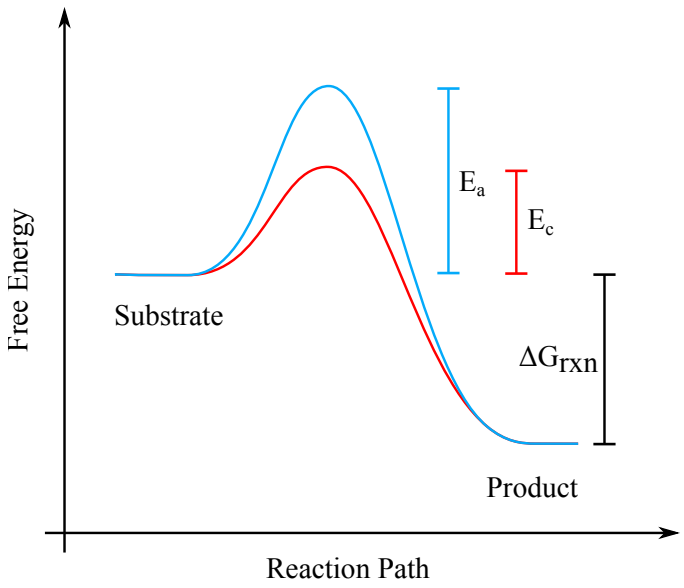

Figure 1.2: Mechanism of catalysis. E_a and E_c are the energies of activation of a uncatalyzed and catalyzed reaction. ΔG_{rxn} is the free energy change of the reaction.

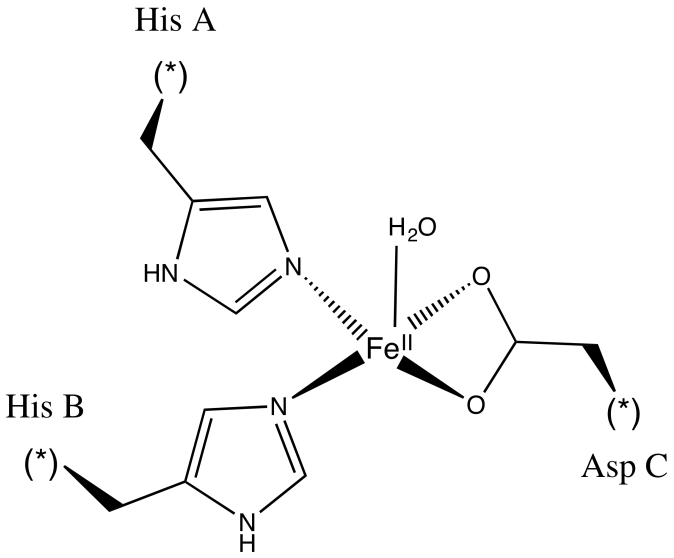

Figure 1.3: General structure of the 2-his-1-carboxylate facial triad along with iron center found in Rieske non-heme dioxygenases.

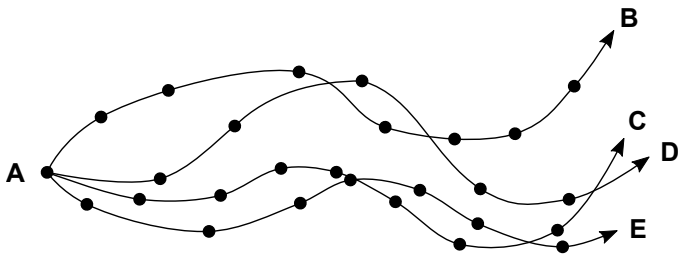

Figure 1.4: Schematic of the enzyme-ligand dissociation process from the bound state **A** to the dissociated states **B**, **C**, **E** and **E** through ligand binding intermediates (marked as circles). Each binding intermediate has entropic and energetic barriers that must be overocome. The optimum trajectory is the one which minimized the barriers. (Adapted from Rydzewski et. al.[18])

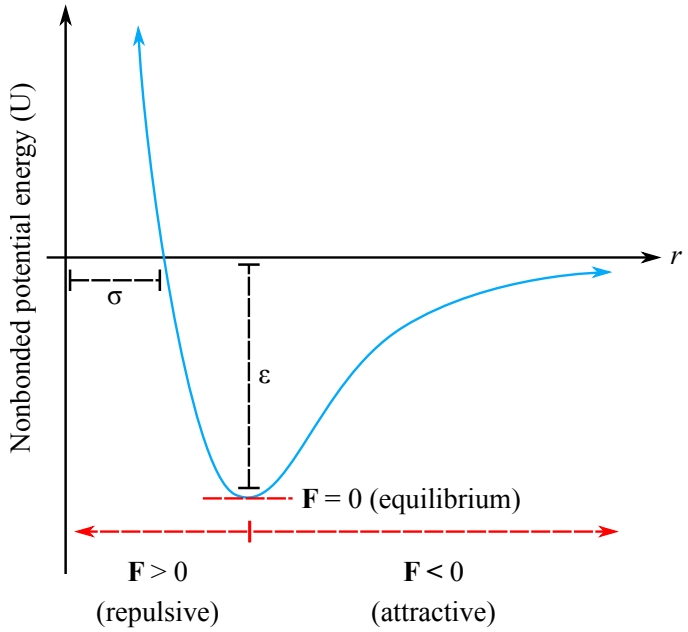

Figure 1.5: Intermolecular potential energy function.

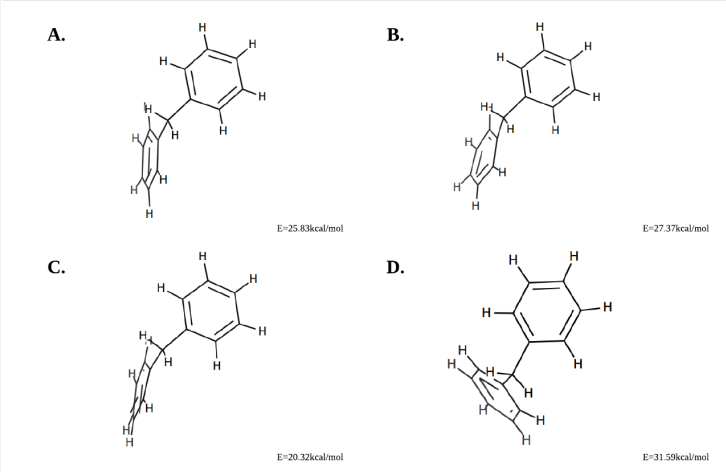

Figure 1.6: A-D Show different geometrical and spatial conformations for Diphenylmethane. This molecule has a flexible bond connecting the two benzyl rings which can rotate, bend and stretch. Some of these conformations are more energetically favorable than others. In order to fit through the tunnel (enzyme selectivity purposes) the molecule might need to pay with a high energy conformation.

Chapter 2

Simulation of the bottleneck controlling access into a Rieske active site: predicting substrates of naphthalene 1, 2-dioxygenase

Adapted with permission from **Escalante, D. E.**, Aukema, K. G., Wackett, L. P., & Aksan, A. (2017). Simulation of the bottleneck controlling access into a Rieske active site: predicting substrates of naphthalene 1, 2-dioxygenase. *Journal of chemical information and modeling*, 57(3), 550-561. doi:10.1021/acs.jcim.6b00469.

2.1 Chapter Summary

Naphthalene 1,2-dioxygenase (NDO) has been computationally understudied despite the extensive experimental knowledge obtained for this enzyme, including numerous crystal structures and over 100 demonstrated substrates. In this study, we have developed a substrate prediction model that moves away from the traditional active site-centric approach to include the energetics of substrate entry into the active site. By comparison with experimental data, the accuracy of the model for predicting substrate oxidation is 92%, with a positive predictive value of 93% and a negative predictive value of 98%. Also, the present analysis has revealed that the amino acid residues that provided the largest energetic barrier for compounds entering the active site are residues F224, L227, P234, and L235. In addition, F224 is proposed to play a role in controlling ligand entrance via $\pi - \pi$ stacking stabilization as well as providing stabilization via T-shaped $\pi - \pi$ interactions once the ligand has reached the active-site cavity. Overall, we present a method capable of being scaled to computationally discover thousands of substrates of NDO, and we present parameters to be used for expanding the prediction method to other members of the Rieske non-heme iron oxygenase family.

2.2 Introduction

Naphthalene 1,2-dioxyegenase (NDO), isolated from *Pseudomonas* sp. NCIB 9816-4, is a Rieske non-heme iron dioxygenase that has been heavily studied with respect to identifying substrates that it will oxidize and determining X-ray structures with substrates bound.[25,27,46,95,96] It is a good example of an enzyme that should be modeled by the keyhole−lock−key model, as it has a 15Å access channel connecting the bulk solvent and the active site.[57] This enzyme alone has been shown to be responsible for the degradation of over 100 different recalcitrant pollutants[7,25,48] that are found in polluted water and soils, making it useful for bioremediation. In addition, the inherent enantio- and regiospecificity of the enzyme makes it very attractive for synthetic applications such as the large scale biosynthesis of the blue jean dye indigo.[46] NDO, like other Rieske dioxygenases, is composed of a three-component system that includes an NADH-dependent flavoprotein reductase, a Rieske [2Fe−2S] ferrodoxin, and the $\alpha_3\beta_3$ terminal oxygenase.[50] During the catalytic cycle, an electron transfer chain is established, transferring electrons one at a time from the reductase to the ferrodoxin and finally to the terminal oxygenase. The catalytic domain of NDO is located in the subunit, where a high-spin mononuclear Fe^{II} is bound to two histidines (H208 and H213) and a bidentate aspartate (D362) that together form the 2-his-1-carboxylate facial triad motif.[50,57]

On the basis of structural observations, Ferraro et al.[21] proposed that the shape and size of the entrance to the active site of NDO (Figure 2.1) might keep

larger substrates out of the NDO active site even though there would be enough space to accommodate those ligands in the active site. They explained this by describing the entrance to the active site as an inverted funnel with an aperture leading to a large vestibule. Despite the large body of information about NDO, including crystallographic data,[19,26,95] single-turnover assays,[57,97] site-directed mutagenesis at active-site residues,[30] and whole-cell assays,[7,25]there is still no fast and accurate way of predicting possible substrates of NDO.

Since there are thousands of possible substrates of NDO, it would be impossible to test them all experimentally. Therefore, a prediction method that is computationally inexpensive and can elucidate the powerful catalytic capabilities of enzymes like NDO is needed. In a recent review, Chen[98] outlined several structure and ligand-based approaches for the computational prediction and identification of enzyme substrates. Chen also proposed that integrated frameworks considering different aspects of the ligand-binding event have to be considered in order to improve prediction accuracy.[98] 3 Therefore, in this work we present a novel hybrid algorithm for the prediction of substrates of NDO that takes into consideration: i) the structural dynamics of the enzyme, ii) the process of ligand entry into the active site, iii) the nonbonded interaction potential between the enzyme and the ligand as it moves along the access channel, and iv) the configuration-space energetics of the ligand once it has reached the active site. Although we are limited by the analysis of only the nonbonded potential along the path, our method can identify high energy barriers and determine whether a chemical compound will be able to overcome any barriers (energetic + geometric) imposed by the channel architecture. In addition, by the use of a multiparameter

approach, it is possible to determine that a chemical has a favorable entry route, the frequency at which it can enter the active site, and whether there exist any favorable interactions inside the active site. Using these parameters, we were able to generate predictions of the likelihood of compounds being substrates of NDO. Since the general structure of the Rieske non-heme iron dioxygenases is conserved throughout the family, a similar method can be applied to study other members of this family of enzymes and generate substrate predictions.

2.3 Methods

2.3.1 Preparation of Crystal Structure and Ligands

All of the calculations were conducted using the Schrödinger molecular modeling suite (version 2015-3). The unbound crystal structure (PDB code 1NDO) was obtained from the Protein Data Bank and prepared using the Protein Preparation Wizard[83] as described elsewhere.[80,82] The only resolved crystal water molecule retained during the model preparation was the one ligated to the mononuclear iron. A total of 45 ligand structures and degradation data were obtained from Aukema et al.[7] and Seo et al.[48] (Figure 2.2, Table 2.1 and 2.2). The structures were prepared using LigPrep as described elsewhere,[80–82] including a full minimization using the OPLS3 force field,[99] Schrödinger's default force field for small molecules.

2.3.2 Changes in Active-Site Conformation

The root-mean-square deviation (RMSD) from the unbound structure (PDB code 1NDO) was calculated for available bound structures of NDO with nonrepeating ligands (Table 2.3). Only the heavy atoms (C, N, O, S) in the 17 residues making the active site[26] were used to determine the RMSD. The process was repeated for biphenyl dioxygenase, another enzyme in the same family, with the bound (PDB code 1ULI) and unbound (PDB code 1ULJ) structures (Table 2.4).

2.3.3 Molecular Dynamics System Building and Simulation

For MD simulations, the prepared unbound structure of NDO was soaked in an orthorhombic water box with a 10 Å buffer between the protein surface and the box walls. The single point charge (SPC) water model was used, and counterions were added to maintain electrical neutrality. The system was relaxed and equilibrated through a series of minimizations and short molecular dynamics simulations using the standard relaxation protocol in Desmond.[100,101] To ensure equilibration of the water molecules in the binding pocket and the bulk, the solvate pocket utility in Desmond was used.[102] Finally, the production simulation was run for 40 ns at a temperature of 300 K and a pressure of 1 atm. The thermostat and barostat conditions, as well as the cutoff radius for long electrostatic interactions, were set as described by Shivakumar et al.[100] The energy and trajectory snapshot coordinates of the simula-

tion were recorded at 5 ps intervals. The convergence of the simulation was checked by plotting the variation in the RMSD as a function of the simulation time.

2.3.4 Tunnel Identification

Using all of the frames in the last 20 ns of the MD trajectory, we identified possible channels leading from the solvent region into the active site using the program (see Appendix A for a sample input script). The starting point for the channel identification was forced to be the non-heme iron, as this is known to be the catalytic site of NDO.[26,57] Hereafter we refer to the location of the non-heme iron as the end of the channel. For each of the 4000 MD snapshots, either none, one, or multiple possible channels connecting the iron to the solvent region were identified. In order to prune the channels, we specified that the correct channel leading from the solvent region to the active site would need to fit the following criteria: i) it must be formed by at least nine active-site residues (50% of the reported number of amino acids forming the active site[26]); ii) it must be at least 15 Å in length, as reported by Wolfe et al.;[57] and iii) it must contain F224 as part of the channel, since this has been reported to be at the entrance opening.[20,21] (These criteria are applicable to NDO, and they would need to be appropriately adjusted to correctly identify channels in other enzymes.) Each of the channels identified by MOLE 2.0 returns a set of coordinates describing the centerline of the path connecting the non-heme iron and the bulk solvent. Hereafter we call each individual point along the path of the channel a step and denote it by the subscript k A total of 100 trajectory snapshots with correct channels were randomly

selected from the pool of 4000 available to sample the pose space and represent the population of possible tunnel conformations.

2.3.5 Docking and Scoring

The structure of NDO with bound naphthalene (PDB code 1O7G) and 100 selected trajectory snapshots were used to generate Glide scoring grids for docking calculations. For each geometry, a docking grid box of default size (20Å × 20Å × 20Å) was centered on the mononuclear iron near the active-site region. Default force field parameters (OPLS_2005[100]) were used, and no additional constraints were defined during the grid generation. The 45 prepared ligands were docked into all of the grids using the Glide docking protocol with default parameters, and GlideXP was selected as the final scoring function. From the docking results, we defined the stopping distance of a ligand (r_μ) as the average distance between its center of mass and the mononuclear iron in all 100 frames. In addition, we defined the distribution distance (r_σ) to be one standard deviation away from the average (r_μ).

2.3.6 Tunnel Traversal and Ensemble Energy

To simulate the movement of the ligand along the channel and evaluate the non-bonded potential energy, we developed an in-house channel traversal algorithm implemented in C++. The basis of this algorithm is sampling of the non-bonded poten-

tial between a ligand and the surrounding enzyme environment at discrete locations (steps) along the access channel. At each step along the channel, the ligand is capable of sampling thousands of different possible orientations by rotating about its center of mass. We simulated this rotation using a spherical coordinate system described by angles θ and ϕ (See Appendix B for more details). To minimize the computational load, the ligand vibrations were ignored, so this analysis considers only the rototranslational effects of the ligand in the direction of the channel. Movement normal to the direction of the channel was not explicitly studied, but we expect that using 100 channel configurations was sufficient to exhaustively sample the ligand pose space relative to the channel. Figure 2.3 outlines all of the preparatory steps (left panel) and gives an outline of the algorithm (right panel). The pseudocode can be found in Appendix C.

The potential energy function is calculated using the widely accepted force field formula:[43]

$$E = E_{\text{bond}} + E_{\text{angle}} + E_{\text{nonbonded}} + E_{\text{torsion}} \tag{2.1}$$

For each snapshot, the unbound geometry is its own reference state, and thus, all of the bond, angle, and torsion terms collapse in the formula. This is a limitation of our method since we do not allow any induced rearrangement of the enzyme or the ligands. From this simplification, the resulting expression to calculate the nonbonded potential is described by eq 2.2:

$$\Delta E_{\text{E-L}}(\mathbf{x}, k) = E_{\text{nonbonded}} \tag{2.2}$$

where $\Delta E_{\text{E-L}}$ is the enzyme−ligand interaction energy at step k. The equivalent Cartesian coordinates described by the rotation about angles θ and ϕ are collectively denoted as $\mathbf{x}$ (See Procedure 3, lines 2-4 in Appendix C for the conversion from spherical to Cartesian coordinates). The enzyme−ligand interaction energy at step k is given by eq 2.3:

$$\Delta E_{\text{E-L}}(\mathbf{x}, k) = \sum_{i}^{\text{on L}} \sum_{j}^{\text{on E}} \left[\frac{q_i q_j e^2}{r_{ij}} + 4\epsilon_{ij} \left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^{6}}{r_{ij}^{6}} \right) \right] \tag{2.3}$$

where subscripts i and j refer to heavy atoms (C, N, O, S) in the ligand and the enzyme, respectively, r_{ij} j is the linear distance between atoms i and j, q_i is the partial charge of atom i, and σ_{ij} and ϵ_{ij} are the short-range interaction parameters for atoms i and j. Standard combination rules ($\sigma_{ij} = (\sigma_i + \sigma_j)/2$ and $\epsilon_{ij} = (\epsilon_i \epsilon_j)^{1/2}$) were applied to obtain interaction parameters.

2.3.7 Assignment of Force Field Parameters

Most of the atoms in our system have been well described and parametrized by a combination of the self-consistent OPLS force fields for amino acids,[43] small organic molecules,[44] molecules containing nitrogen and oxygen,[45] and halogenated molecules. The values used to describe the partial charges q_i and q_j and short-range interaction parameters σ_{ij} and ϵ_{ij} in eq 2.3 for all of the ligand and enzyme atoms were obtained from the OPLS force fields for amino acids[43] and small molecules.[44,45,103] For the mononuclear +2 iron center, we determined the partial charge using an effective

core potential as described by Bassan et al.[42,51] and then fitted the electrostatic potential using the Merz−Singh−Kollman scheme, similar to the protocol described by Jambeck et al.[104] When using this protocol, we employed density functional theory with the B3LYP exchange−correlation functional and the 6-311+G(d,p) basis set. In order to account for polarization effects inside the highly hydrophobic cavity, we used a lower dielectric constant ($\epsilon = 4$) as justified by Bassan et al.[42,51] For the short-range electrostatic interactions (Lennard-Jones), we used the previously derived parameters $\sigma_{Fe^{2+}} = 1.386$ and $\epsilon_{Fe^{2+}} = 0.0136$.[105]

2.3.8 Calculation of Potential and Channel Continuity

For every step k along the channel, we used a 51×26 grid ($-\pi < \theta \leq \pi, -\pi/2 < \phi \leq \pi/2$) to obtain a total of 1326 orientations. Each of these orientations can be described as a microstate that the ligand can attempt to access. For each of these orientations at step k, the nonbonded potential energy was evaluated using 2.3 (i.e., we evaluated a total of $1326 \times k$ interactions for every ligand-frame combination). The probability of accessing each of these microstates can be determined using the Gaussian probability $\rho(\mathbf{x}, k)$ given by eq 2.4

$$\rho(\mathbf{x}, k) = \frac{e^{-\beta \Delta E_{\text{E-L}}(\mathbf{x},k)}}{\int e^{-\beta \Delta E_{\text{E-L}}(\mathbf{x},k)} d\mathbf{x}} \tag{2.4}$$

where $\beta = 1/\text{RT}$

The configurational entropy contribution, S_{config}, was calculated using eq 2.5:

$$S_{\text{config}} = -R \sum_{\mathbf{x} \in M} \rho(\mathbf{x}, k) \ln \rho(\mathbf{x}, k) \tag{2.5}$$

where M is the set of all possible microstates that are accessible. For each microstate at step k, we calculated the free energy $\Delta G(\mathbf{x}, k)$ using eq 2.6:

$$\Delta G(\mathbf{x}, k) = \Delta E_{\text{E-L}}(\mathbf{x}, k) - T S_{\text{config}}(k) \tag{2.6}$$

The calculated nonbonded free energies of all microstates $\mathbf{x}$ in step k, $\Delta G(\mathbf{x}, k)$, were used to build nonbonded potential energy maps (E-maps) for each step k along the channel. We used the E-maps to determine the channel continuity (CC), a Boolean parameter based on the condition defined in Figure 2.3. The CC was determined for each ligand−frame combination, and the average number of times that the CC condition was met for each ligand (i.e., the total number of times that the ligand reached the active site) was defined as the percentage entrance rate. Finally, we used eq 2.7 to calculate the ensemble (Gaussian) average free energy at step k denoted as $\langle \Delta G(\mathbf{x}, k) \rangle$:

$$\langle \Delta G(\mathbf{x}, k) \rangle = \frac{\int \Delta G(\mathbf{x}, k) e^{-\beta \Delta G(\mathbf{x}, k)} d\mathbf{x}}{\int e^{-\beta \Delta G(\mathbf{x}, k)} d\mathbf{x}} \tag{2.7}$$

We then used the calculated $\langle \Delta G(\mathbf{x}, k) \rangle$ values for all k steps along the channel to build a nonbonded potential trajectory profile specific to each ligand−frame combination. The trajectory profile calculation was repeated for all 4500 ligand− frame combinations (45 ligands × 100 frames). Finally, from the 100 trajectories for each

ligand, we calculated the (arithmetic) average trajectory profile for each ligand (Figure C.3), resulting in a total of 45 average trajectory profiles. The average trajectory profile was used to calculate the average free energy of interaction between the enzyme and ligand as it enters the active site (ΔG_{trj}) and the average free energy of interaction inside the active site ($\Delta G_{\mathrm{A.S.}}$). Further details describing the calculation of the free energies can be found in Appendix C.3 and Figure C.3.

2.4 Results and Disucission

2.4.1 Substrate Prediction via Stand-Alone Docking Methods

In preliminary studies, molecular docking was utilized in an attempt to predict ligands using a single-crystal structure of NDO (PDB code 1O7G). We used the Glide program[81] with the GlideXP scoring function to dock our small library of 45 known NDO substrates and nonsubstrates (Tables 2.1 and 2.2).[7,106] The initial screening scored all of the compounds within a very narrow docking score range of 1.8 kcal/mol (data not shown). The narrow energy range did not allow us to set a cutoff value to differentiate between substrates and nonsubstrates of NDO.

It was previously reported that one of the shortcomings of docking as a stand-alone predictive method is that it uses a single static conformation of the enzyme[58]

obtained from a crystal structure or homology modeling. This poses a great challenge since it is known that proteins breathe and fluctuate between different microstates within the same conformation and assume different macrostates with distinctly different conformations when they are hydrated.[68] Therefore, it was not surprising that our preliminary docking analysis with the single static structure of NDO did not fully capture the true binding capability of the active site. To model the enzyme breathing motion and in an attempt to obtain a better scoring distribution, we used an approach suggested by the original authors of the GlideXP scoring function, which required docking the ligands in different protein conformations of the same receptor.[80] We used 100 different conformations of NDO obtained from a molecular dynamics simulation. The range of docking scores increased to 4 kcal/mol, as shown in Figure C.4, but the distribution of the scores we obtained still did not allow us to define a cutoff score to discriminate between substrates and nonsubstrates of NDO. We observed that the predictive capabilities of the docking method for NDO were insufficient for substrate prediction, which was consistent with results reviewed by Chen[98] for the prediction of substrates of cytochrome P450.

Another problem we identified when using docking as a stand-alone method for prediction was that very large nonsubstrates of NDO, such as 9,10-dihydro- 9,10-methanoanthracene (compound **4**), reported lower docking scores (i.e., more favorable interactions) than smaller known substrates, such as naphthalene. Since the GlideXP scoring equation uses over 80 empirically calculated parameters to assign rewards and penalties,[80,81] it is very likely that these parameters influenced the way that Glide calculated the final docking score and assigned lower scores for the larger molecules.

Furthermore, the lack of prediction accuracy observed from our results indicated that NDO likely falls outside the applicability domain of the GlideXP reward and penalty parameters. However, ignoring the path that a molecule has to traverse to reach the active site is a significant oversimplification that can reduce the predictive power of docking as a stand-alone method. Specifically for an enzyme such as NDO, with a geometrically nonspecific active site, the docking program fits large ligands inside the cavity without regard for any physical hindrances that occlude access to the active site.

On the basis of our docking results and the results reviewed by Chen,[98] more advanced computational methods to supplement docking are necessary. SMD is a method that can be used to supplement docking as a prediction method. In this approach, it is possible to investigate the energetic barriers that a ligand faces as it enters or exits the active site of an enzyme,[107] and this method has been successfully used to investigate how substrates enter and products exit the buried active site of cytochrome $P450_{cam}$.[16] However the main hurdles for using this method are i) the computational time that it requires to complete all the necessary simulations and ii) the fact that the biased external forces applied to pull the ligand out of the active site require prior knowledge of the final conformational state of interest.[108] Since our main goal in this study was to develop a fast and accurate prediction method capable of assessing a large number of compounds that could be extended to other enzyme systems, SMD was considered to be prohibitively expensive computationally. Therefore, in this study we proposed a hybrid simulation method that allowed us to develop a faster prediction method than MD or SMD simulations yet retain an analysis

of the dynamic nature of the enzyme. For this method we assumed that NDO is a "stiff" enzyme whose backbone does not experience major permanent configurational changes in the active site (i.e., domain motions) upon ligand binding and that all of the conformational changes are due to side-chain movements explorable within medium length MD simulations (10ns).

In order to test the validity of our "stiff" backbone assumption, we calculated the RMSD of the $\alpha-$carbons in the active site for the available holo (bound) structures with nonrepeating ligands from the apo (unbound) structure. The average RMSD was calculated to be 0.57 Å (see 2.3 for the individual values). We compared our value to the RMSD of 5.76 Å for the bound structure of biphenyl dioxygenase (BPDO) from *Rhodococcus* sp. RHA1 (2.4). This variant of BPDO has been reported to show a significant rearrangement of the active-site backbone upon binding of biphenyl compared with the lack of rearrangement observed in NDO upon binding of the same ligand.[109] This supported the hypothesis that the backbone of NDO is in fact "stiff" and does not undergo major permanent structural rearrangements upon ligand binding.

2.4.2 Molecular Dynamics Simulations and Tunnel Identification

We performed a single 40 ns MD simulation with an empty active site in order to obtain a range of possible conformational states that the enzyme might adopt over

time. A caveat of simulating an empty active site is that it does not identify any possible transient conformational changes as large ligands make their way into the buried active site, in the way that SMD or REMD simulations would. However, our approach is rapid and can allow predictions to be made for 10^5 possible different ligands in the same amount of time as a single comprehensive SMD simulation.

We analyzed the $\alpha-$carbons RMSD and root-mean-square fluctuations over the 40 ns simulation time to determine the range of motion of the NDO enzyme within the MD simulation. We observed that the simulation equilibrated approximately 20 ns after the start of the simulation. The maximum deviation from the initial structure reached approximately 3.4 Å with an average RMSD for all heavy atoms of 3.0 Å, as shown in Figure 2.4. In addition, the average RMSD for the 17 binding pocket residues[26] and the mononuclear iron center was 0.65 Å, which supported our earlier observation that there were no major deviations in the backbone structure from that of the original X-ray structure. In addition, the RMSF analysis over the last 20 ns of the simulation (data not shown) demonstrated that the motions exhibited by the individual amino acids throughout the entire enzyme do not undergo major changes. The only amino acids that showed major fluctuations were at the N- and C-termini.

We analyzed all of the simulation frames between 20 and 40 ns (a total of 4000 frames) to identify all possible channels connecting the buried active site and the solvent region. As expected, all of the simulation snapshots showed channels with different geometric characteristics. Although the backbone of NDO is considered stiff, its side chains still experience a considerable degree of motion within the constraints

imposed by the position of the backbone. From our tunnel identification analysis, we found that only 70% of the snapshots had an open channel configuration, while the rest of the snapshots showed no connection between the active site and the solvent region. This fluctuation between open and closed channels can be attributed to the movement of specific amino acid side chains. Finally, from the full simulation trajectory analysis we determined that the average length of the tunnel was 17 ± 2Å, which can be compared to the previously reported length of 15 Å determined from a single static structure.[57]

We randomly selected 100 snapshots that showed an open configuration (3.6% of the snapshots with an open channel population) to represent most of the different geometries and dynamics that the channel adopted during the MD simulation. The average channel radius of the sample, calculated and defined by MOLE 2.0, showed that the tunnel had a wide entrance and then narrowed, forming a bottleneck $10-15$Å away from the iron (Figure 2.5a). We observed that the active site cavity was large ($\sim$ 10Ålength$\times$ $\sim$ 5Ådiamter) with the mononuclear iron site located opposite to the channel entry. These results are consistent with the description of the entrance into the active site provided by Ferraro et al.[21] such that the shape is an inverted funnel leading to a large vestibule.

Next, we analyzed the wall lining properties of the 100 open channel frames as determined using MOLE 2.0. We found that a total of 32 different residues formed the walls of the channel. The frequency with which these residues contributed to forming the wall varied from 40 to 100% of analyzed frames. We calculated the distances

between the mononuclear iron and the center of mass of each of the 32 amino acids in all 100 frames and plotted them as a heat map (Figure 2.5b). We showed only the amino acids that contributed to forming the wall in at least 80% of the analyzed frames. Given the identified location of the bottleneck $10 - 15$Å away from the from the end of the tunnel (Figure 2.5a), we determined from the heat map (Figure 2.5b) that the amino acids most frequently found in the $10 - 15$Å distance range were F224, L227, P234, and L253 (Figure 2.5c). F224 in the entrance channel of NDO is bulkier than the corresponding L223 in BPDO$_{B1}$,[21] and thus, our identification of the bottleneck residues was consistent with the proposal by Ferraro et al.[21] that the shape and size of the entrance of NDO may be a factor keeping larger substrates out of its active site.

2.4.3 Analysis of the Active-Site Solvation State

In order to further simplify our model and decrease the computational load of our in-house algorithm, we assumed that solvation could be neglected. The validity of this assumption depended principally upon the free energy to be gained by displacing the water molecules[80] along the trajectory toward the active site. To verify that our assumption to ignore solvation effects was valid, we analyzed the positions of all of the water molecules within 10 Å of the channel centerline (Figure 2.9) in all 100 selected frames. We found that all of the water molecules were located within 3.6 Å of the channel centerline (Table 2.7). The smallest molecule we analyzed was benzene (3.4 Å in height based on the van der Waals radius of carbon), so it was reasonable to

assume that any orientation of benzene displaced all of the water molecules along the path. By extension, wider ligands would also displace all of the water molecules as they entered into the active site of NDO. Since all of the water molecules would be equally displaced by all of the ligands, the net relative solvation effect would be zero. We are aware that this simplification was valid when studying NDO but may not apply for other enzymes of the same family, such as biphenyl dioxygenase or toluene 2,3-dioxygenase (TDO), for which the distributions of water molecules inside the channel are different because of the differences in size and hydrophobicity of the active site (data not shown).

2.4.4 Assignment of Fe^{II} Partial Charge

As our final preparatory step, we calculated the partial charge of the Fe^{II} to be 1.304 based on the methods proposed by Bassan et al.[42,51] and Jambeck et al.[104] We used a simplified model of the 2-His-1-carboxylate motif with the mononuclear iron formal charge of +2 (i.e. Fe^{II}) as shown in Figure 2.6. We chose this simplified version since it was demonstrated that small cluster models can provide accurate results and reduce the computational load.[110] The oxidation state of iron was defined to be +2 since in all of the proposed mechanisms (based on experimental[57,97,111] and computational studies[42,51]) the mononuclear iron changes oxidation state from the the Fe^{II} resting state to the Fe^{III} state upon binding of oxygen (one step after binding of the ligand). We understand that the value we have calculated is valid only for nonbonded calculations like the ones used in our simplified model. A full

quantum study as described elsewhere[112,113] would be necessary to fully parametrize the 2-His-1-carboxylate facial triad motif.

2.4.5 Channel Electrostatic Mapping and Path Continuity

The model that we developed accounted for geometric and nonbonded energetic effects imposed by the NDO enzyme channel on the ligand during its trajectory into the active site. We used the output of MOLE 2.0 to determine the centerline of the channel along the path connecting the two ends (the active site and bulk solvent). This centerline was divided into k steps of variable size in the range $0.05 - 0.15$Å; ; this allowed us to move each ligand along 170−225 steps depending on the length of the channel at each snapshot.

We performed our roto-translational simulations (as outlined in the right panel of Figure 2.3) for all enzyme−ligand combinations. We used the results from our simulations to calculate the nonbonded potential energy map (E-maps) for all enzyme−ligand−step combinations. An example E-map for naphthalene located 12.1Å from the iron is shown in Figure 2.7a (the E-maps for the other 4499 enzyme−ligand combinations are not shown). There were some microstate clusters that had negative free energy ($\Delta G < 0$kcal/mol), indicating that the ligand at those given orientations had a favorable interactions with the enzyme. Other microstate orientations were found to be energetically unfavorable ($\Delta G > 0$kcal/mol) Finally, we found that in many unfavorable microstates, one or more atoms of the ligand overlapped the enzyme

(these orientations showed $\Delta G > 600$kcal/mol). Since two atoms cannot occupy the same space (overlap), these microstates were said to be inaccessible and are shown as the white regions in Figure 2.7a.

We used all of the E-maps for a given ligand−frame combination to calculate the free energy along the ligand trajectory into the active site (Figure 2.7b). The energy maps also allowed us to calculate a new parameter that we have termed the *channel continuity* (CC). For a ligand to successfully enter the active site, contiguous accessible microstates must exist for each set of nearest-neighbor steps.The CC condition is shown in Figure 2.3 as $\Delta G(\mathbf{x}, k-1) \cap \Delta G(\mathbf{x}, k) \cap \Delta G(\mathbf{x}, k+1) \neq \varnothing$. Both of the ligands shown in Figure 2.7b had at least one accessible microstate in every step k; however, the red $\times$ symbol for each indicates the step at which the channel continuity between accessible microstates ceased to exist.

From the profile of the trajectory free energy (Figure 2.7b), we were able to predict that bulkier molecules, such as adamantane, were less likely to be substrates of NDO *in part* because they clashed with the channel walls in the bottleneck region and did not meet the channel continuity criterion. On the other hand, substrates of NDO such as naphthalene had more favorable energetic profiles as well as channel continuity all the way into the active site. Having calculated the trajectory profiles for a given compound (Figure C.1 left), we then evaluated the average energy profile over the analyzed frames (Figure C.1 right). The trajectory profiles shown in C.2 represent the average nonbonded potential energies between the ligand and enzyme for the selected 100 frames. These average trajectory profiles allowed us to calculate

ΔG_{trj} and $\Delta G_{\text{A.S.}}$ in order to develop a better substrate prediction model.

2.4.6 Multiparameter Prediction Approach

For NDO, three parameters were necessary to adequately model both the access-channel keyhole and the active-site lock. The first parameter was the channel continuity. This parameter allowed us to determine whether a compound was able to enter the active-site cavity, since entrance is required for catalysis to occur. Given the changing nature of the channel geometry, we calculated an average probability of entry into the active site for each ligand. We found that benzene (known to be a substrate) was able to enter in 98% of the analyzed frames, whereas triphenylene (known to be a nonsubstrate) was not able to enter the active site in any of the analyzed frames (i.e., 0% entrance). A full list of probability percentages for all of the tested compounds can be found in Tables 2.5 and 2.6. Since triphenylene was not able to enter the enzyme at all, it was classified as a nonsubstrate and not considered for further analysis.

The second parameter used to determine the likelihood for a ligand to be a substrate was the average free energy along the accessible trajectory, which we denote as ΔG_{trj} in Figure 2.8 (see Appendix C §C.3 for the calculation of ΔG_{trj}). This was the average nonbonded interaction energy that the ligand experienced with the enzyme at each point along the channel until it reached the active site ($r_\mu + r_\sigma$). Therefore, for a ligand to be a substrate, it had to predominantly have a negative

ΔG_{trj}. On the other hand, the compounds that were less likely to be substrates were blocked by highly unfavorable interactions, and thus, a positive ΔG_{trj} was observed. We understand that the unfavorable interactions we calculated might be higher than the experimental values since we did not allow any induced fitting of the enzyme, as that would increase the computational load for prediction.

The third parameter was the average free energy attainable within the active-site region ($r_\mu \pm r_\sigma$) denoted as $\Delta G_{\text{A.S.}}$ in Figure 2.8. This was analogous to a docking score, except that we only measured the nonbonded interactions and ignored any of the empirically assigned rewards or penalties commonly used by docking scoring functions.[80–82] Similarly, it was expected that compounds able to enter the active site and react would have a negative $\Delta G_{\text{A.S.}}$. It is important to note that the reference state used to calculate ΔG_{trj} and $\Delta G_{\text{A.S.}}$ was the energy of interaction between the enzyme and the ligand at an infinite distance (i.e. G_∞) which is effectively 0 kcal/mol.

2.4.7 Accuracy of the Model and Its Prediction Capability

In order to test our hypothesis and the validity of our hybrid multiparameter prediction approach, we plotted ΔG_{trj} versus $\Delta G_{\text{A.S.}}$ (Figure 2.8) for 45 compounds previously tested experimentally for reactivity with NDO.[7,48] The sample was distributed as follows: very good substrates ($n = 23$), good substrates ($n = 5$), bad substrates ($n = 6$), very bad substrates ($n = 10$), and known inhibitors ($n = 1$), based on the percentage of compound removed from a sealed resting cell assay (Tables 2.1 and

2.2). By using the natural cutoff between positive and negative values, we split the plot into four different quadrants (QI−QIV). We predicted that a successful catalytic event in NDO would be the result of a combination of favorable entrance energetics and favorable interactions inside the active site. Therefore, on the basis of these two parameters, compounds falling in QI were predicted to be substrates of NDO. Our prediction algorithm was in good agreement with experimental data.

All of the compounds found in QI were shown to be either very good or good substrates of NDO.[7]. Similarly, most of the bad and very bad substrates of NDO did not have favorable free energies of interaction along the channel trajectory and were found to be in quadrants II and III. Overall, our prediction accuracy was 92%, with a positive prediction value (true positives) of 93% and a negative prediction value (true negatives) of 98%. To our knowledge, this is the first computational method that has been able to predict with high accuracy the substrates and nonsubstrates of naphthalene dioxygenase or any other Rieske non-heme iron type of oxygenase.

It is worth noting that the entrance favorability (ΔG_{trj}) did not correlate directly with the size of the ligand *alone*. For instance, we observed that the isomers azulene (**a**) and naphthalene (**b**) had very different (ΔG_{trj}) values. The wider compound, azulene, had a 30 kcal/mol more favorable (ΔG_{trj}) value than naphthalene (Figure 2.8 and Tables 2.1 and 2.2). We also found that there was no correlation between (ΔG_{trj}) and the molecular weight or the compound's widest dimension (data not shown). This suggested that entry to the active site of NDO is not merely a size exclusion process but rather is a complex process that depends upon the size,

shape, and distribution of the partial charges in the ligand atoms. From visual inspection of the most energetically favorable configuration of the ligand at the enzyme bottleneck and on the basis of the geometry parameters proposed by McGaughey et al.,[114] we identified the possibility that $\pi - \pi$ stacking interactions form between the aromatic compounds in QI and F224 (as illustrated in Figure 2.10. Thus, one of the several possible factors stabilizing the entrance energetics is $\pi - \pi$ stacking between the aromatic rings of the tested ligand and the aromatic ring of F224. On the other hand, when we studied the interactions of aromatic compounds in QIII (e.g., pyrene (**c**)) with F224, we did not observe any favorable geometry arrangements that would allow the formation of $\pi - \pi$ interactions. One possible explanation for this is that the centroid of any of the four aromatic rings in pyrene is too far away from the centroid of the aromatic ring of F224 (> 3.4Å), as shown in Figure 2.10b.

The true power of this two-parameter approach lies in the ability to discern which of the two parameters (negative energetics inside the active site or along the entrance trajectory into the active site) prevented a compound from being transformed by NDO. This point can be best illustrated by a closer look at the flavonoid compounds. Flavone (**d**) and isoflavone (**e**), in QIV of Figure 2.8, were shown to have catalytic activity in NDO, albeit poor, while (*S*)-flavanone (**f**) and (*R*)- flavanone (**g**) were essentially nonsubstrates.[48,106] Our results suggest that the reason for poor catalytic activity with flavone and isoflavone is due not to their failure to enter the active site but rather to poor conformational placement inside the active site. This is consistent with the results of Seo et al.[48] Despite the close structural resemblance of the two compounds, they were found to fit differently inside the active site. The A

ring of isoflavone (following the nomenclature of Seo et al.[48]) was found to be capable of forming a $\pi - \pi$ stacking interaction with the aromatic ring of F224 (Figure 2.10) in more than 90% of our analyzed frames, an effect not observed when we analyzed flavone for the same type of interaction. This active-site stabilizing effect caused by the T-shaped $\pi - \pi$ interaction is a possible explanation for why ($\Delta G_{\text{A.S.}}$) was much lower for isoflavone than for flavone. On the other hand, (S)-flavanone and (R)-flavanone were very bad substrates of NDO, which on the basis of our predictions is due to their failure to favorably enter the active site of NDO.

Finally, we expect that the power of locating ligands in a quadrant system based on (ΔG_{trj}) versus ($\Delta G_{\text{A.S.}}$) enables possible protein engineering pathways to rationally expand the substrate range of NDO. For instance, to improve the probability that ligands placed in QII are substrates, we propose that the bottleneck (residues F224 or L227) could be mutated in order to allow compounds to have a favorable trajectory into the active site. Similarly, for compounds found in QIV we propose that modifications to active-site residues, like those studied by Yu et al.,[30] would better accommodate ligands inside the active site.

2.5 Conclusions

We have presented a new method to predict substrates of NDO that showed 92% accuracy. In order for our model to predict substrates of NDO at a high accuracy rate and still be computationally inexpensive, we applied the following conditions: i) on

the basis of crystallographic evidence, we ignored any induced-fitting changes in the enzyme as the ligand binds in the active site; ii) we assumed that NDO is stiff and that all of the major conformational changes were observed within a medium-length MD simulation; iii) we assumed that the active-site cavity of NDO is highly hydrophobic, allowing us to ignore solvation effects; and iv) we ignored any vibrational effects of the ligand and considered only roto-translational effects. We are aware that applying this set of conditions and assumptions may be valid only for NDO and that modeling of other Rieske dioxygenases may require the incorporation of more parameters. The results obtained from this study were consistent with previous direct experimental observations of substrates, thus increasing our confidence in the predictive results. This method (besides the initial MD simulation) was not computationally expensive and could be therefore scaled to analyze thousands of chemical compounds and determine their potential as substrates of NDO. Finally, the results presented here open the possibility for the development of prediction methods for other enzymes in the Rieske non-heme iron dioxygenase family as well as possible protein engineering routes to explore and expand the knowledge base of substrates in this class of enzymes.

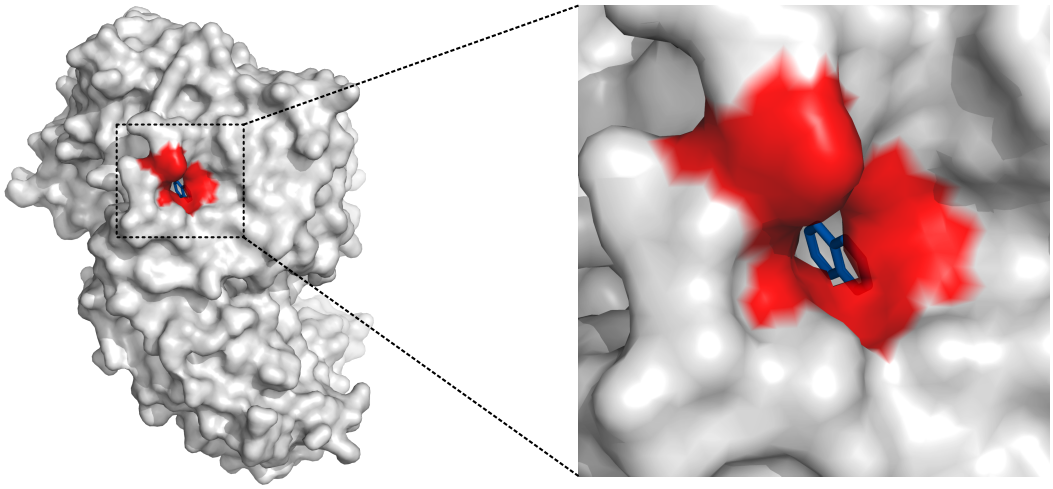

Figure 2.1: General structure of the naphthalene dioxygenase monomer unit from *Pseudomonas* sp. NCIB 9816-4 (PDB code 1O7G). The red region highlights the location where ligands enter the active-site cavity of the enzyme. In the zoomed-in representation on the right, naphthalene (the preferred substrate of NDO) can be seen inside the active site through the channel opening.

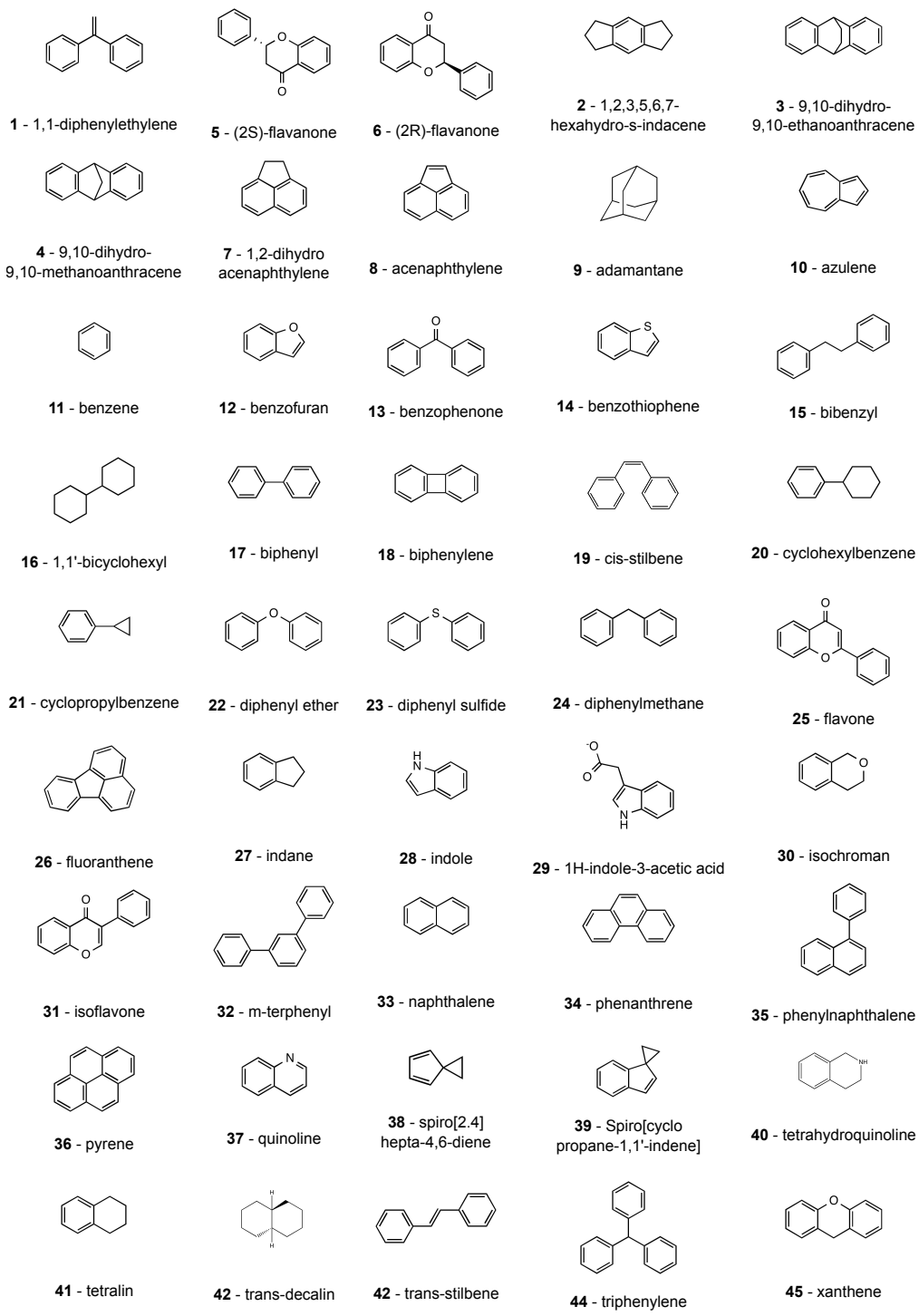

Figure 2.2: Structure of all compounds used in the validation of the all-atom model.

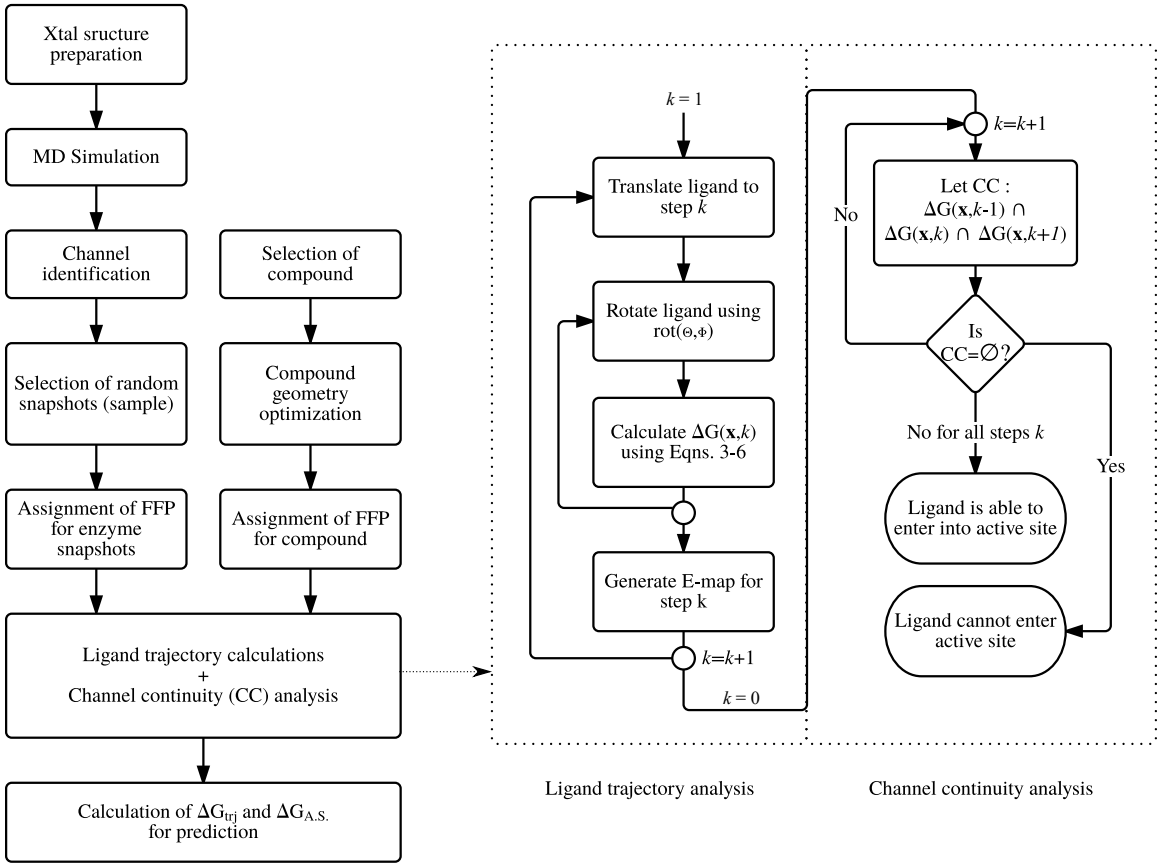

Figure 2.3: (left) Flowchart outlining the preparatory steps and (right) our developed algorithm. FFP stands for force field parameters, and the symbol ∩ represents the data set at the intersection of E-maps.

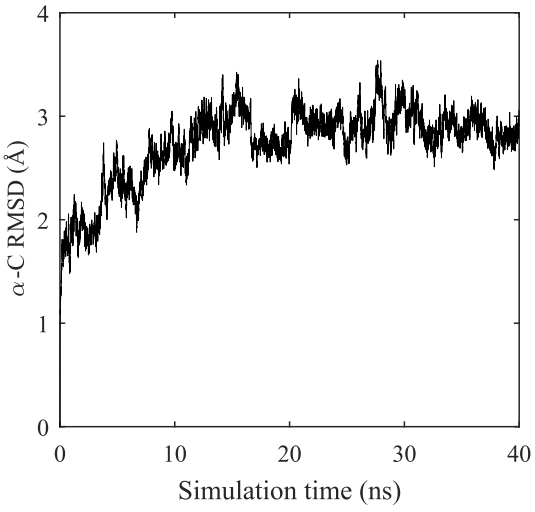

Figure 2.4: Trajectory analysis of the molecular dynamics simulation. The system equilibrated 20 ns after the start of the simulation. After equilibration, the average RMSD for all of the $\alpha-$C atoms was 2.98 Å.

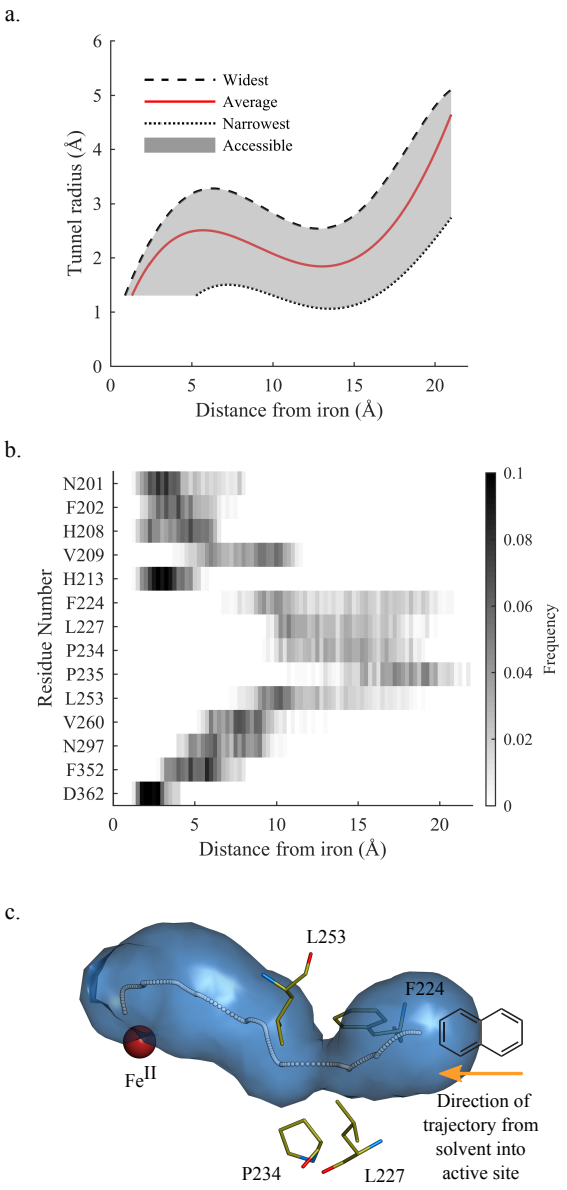

Figure 2.5: Naphthalene dioxygenase channel properties. (a) Radius profile along the NDO channel. The red line shows the average radius for the 100 randomly chosen open configuration tunnels. The upper and lower limits for the channel radius are shown as dashed and dotted lines, respectively. (b) Heat map showing the distances of the centers of mass of the identified wall-forming amino acids from the mononuclear iron. (c) Cartoon diagram showing the channel wall (blue), the four residues that form the bottleneck (sticks), and the centerline of the channel (white dots).

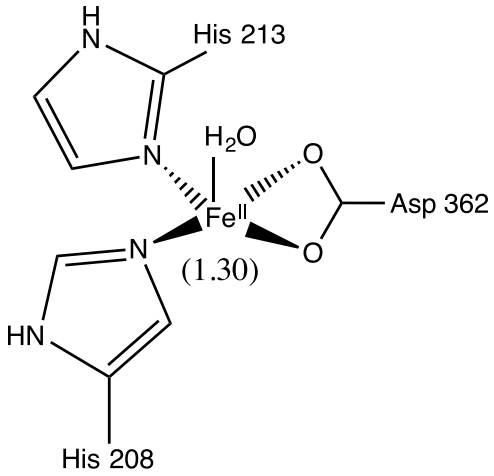

Figure 2.6: Small cluster model of Fe^{II} bound to the two-His-onecarboxylate facial triad used to calculate the partial charge of the mononuclear iron. The calculated partial charge is shown in parentheses under the iron center.

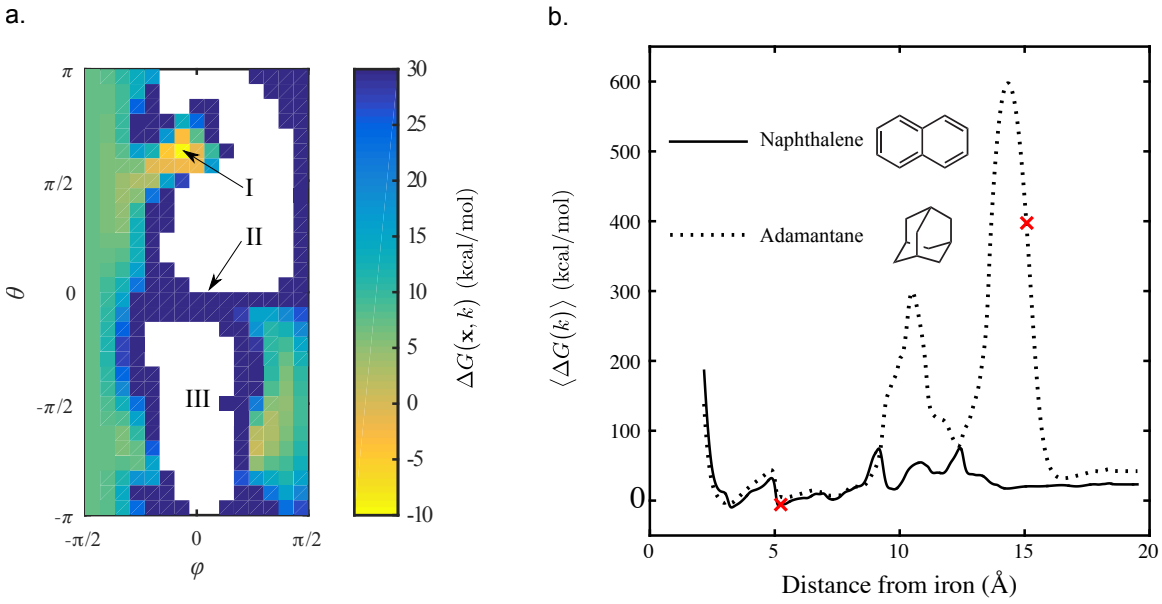

Figure 2.7: Nonbonded potential energy maps (E-maps), free energy of trajectory profiles, and channel continuity. (a) Sample E-map showing the 1326 microstate energy levels of a ligand when rotated by angles θ and ϕ at a single step k along the centerline path. Point I corresponds to the most favorable microstate, and point II shows the least favorable (yet still accessible) microstate. Region III (all white space in the map) represents microstates that are inaccessible because of high-energy interactions resulting from very close proximity between the ligand and enzyme residues. (b) Ensemble-averaged interaction energies along the channel for naphthalene (solid) and adamantane (dotted). The red $\times$ symbols mark the end of the channel continuity. For adamantane, the continuity breaks down at 16.4 Å, whereas naphthalene stops showing continuity at 5.1 Å.

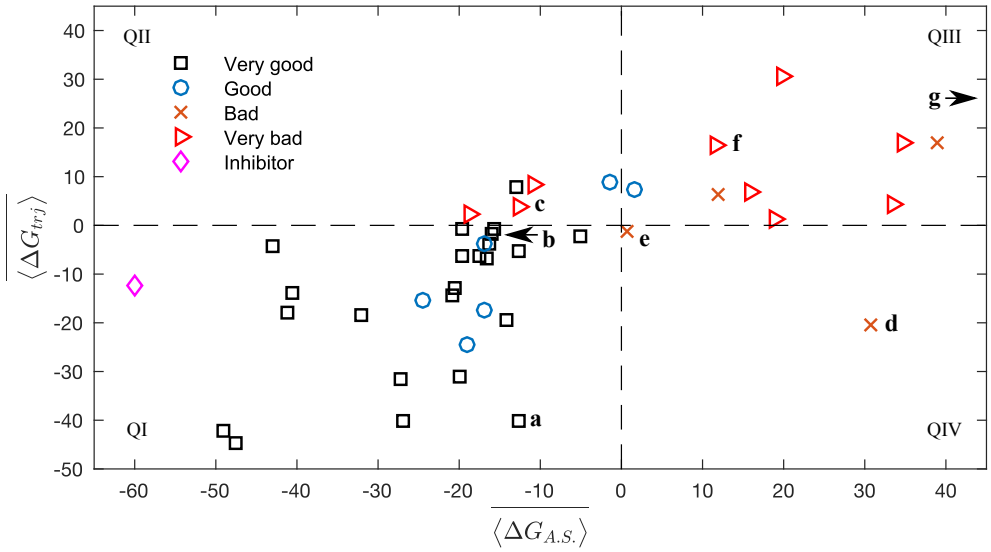

Figure 2.8: Test of our prediction algorithm using set of 45 compounds for which experimental data existed. Very good substrates were defined to be those compounds whose initial concentration decreased $> 75\%$ in a resting cell assay, good substrates fell within the range of 51−75%, bad substrates fell within the degradation range of 25−50%, and very bad substrates were degraded $< 25\%$. The free energy of the trajectory is plotted against the free energy inside the active site. The plot is divided into four quadrants (QI−QIV) labeled in the clockwise direction. (R)-Flavanone (a bad substrate) and *m*-terphenyl (a very bad substrate) that fell in QIII are not shown since $\Delta G_{\mathrm{A.S}} > 40$kcal/mol, Tables 2.5 and 2.6 reports the raw data.

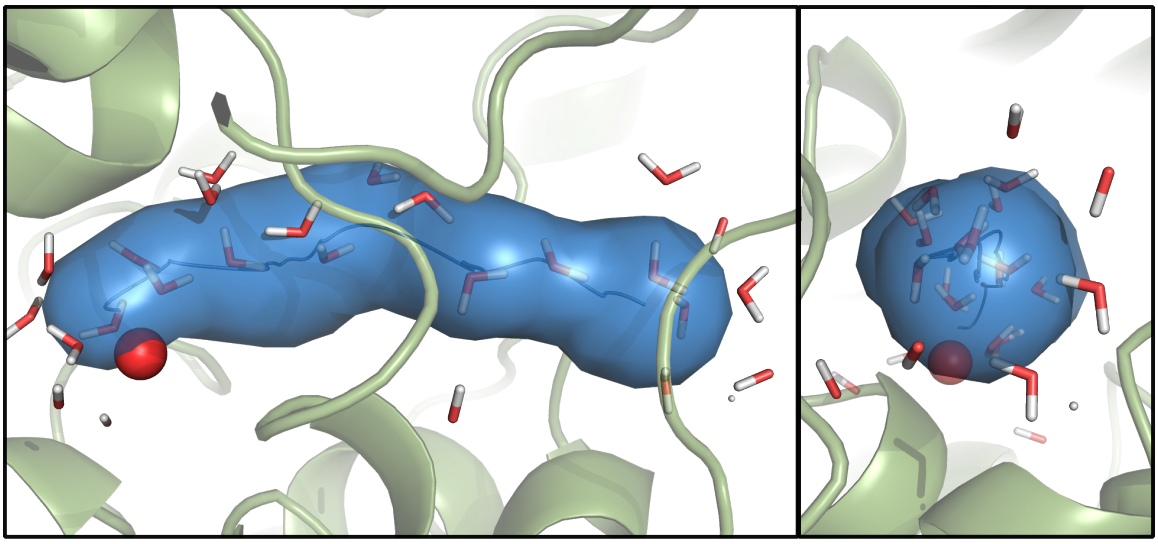

Figure 2.9: Snapshot taken from MD simulation showing all water molecules < 10Å from the channel centerline. The left panel shows a side view and the right panel shows a front view of the same snapshot. The water molecules are shown as stick figures, the mononuclear iron as a red sphere, and the channel centerline and walls (as obtained from MOLE 2.0) is shown as a blue line and surface respectively The enzyme ribbons are shown in green.

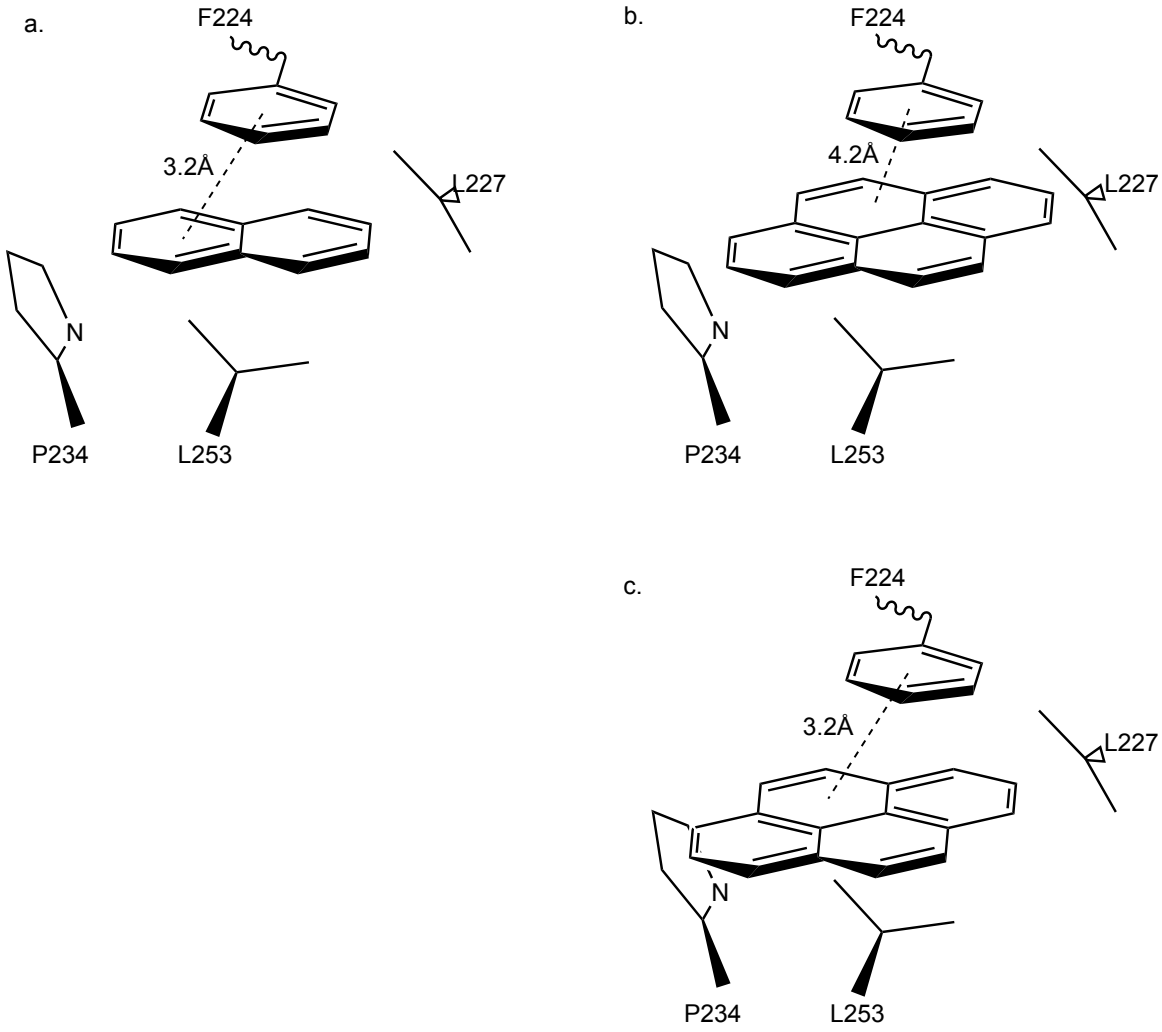

Figure 2.10: Sketch showing the stacked shaped $\pi - \pi$ interaction between F224 and naphthalene and pyrene. a. The distance between the centroid of ring 1 in naphthalene is 3.4 Å away from the centroid of the aromatic ring in F224. b. In the same frame and step along the channel, the closest centroid of any ring in pyrene to F224 is 4.2 Å, too far away to induce stabilization by $\pi - \pi$ interaction. c. The centroid of ring 1 in pyrene has been forced to be at 3.2 Å (overlay to naphthalene in part a) in order to induce a similar stacked shaped $\pi - \pi$ interaction. However, any benefit from this type of interaction is lost due to the overlap of atoms in ring 3 of pyrene and P234.

Table 2.1: Percent removal of compounds in whole cell experiments. Category assignment is as follows: very good $(100 - 76\%)$, good $(75 - 51\%)$, bad $(50 - 26\%)$, very bad $(25 - 0\%)$. Percent removal was calculated based on extremes of standard deviations.

Compound Number	% Removal	Category
1	82	Very good[7]
2	35	Good[7]
3	0	Very bad[7]
4	0	Very bad[7]
5	0	Very bad[48]
6	0	Very bad[48]
7	100	Very good[7]
8	100	Very good[7]
9	0	Very bad[7]
10	100	Very good[7]
11	100	Very good[7]
12	100	Very good[7]
13	41	Bad[7]
14	100	Very good[7]
15	94	Very good[7]
16	0	Very bad[7]
17	100	Very good[7]
18	100	Very good[7]
19	0	Very bad[7]
20	90	Very good[7]

Table 2.2: Percent removal of compounds in whole cell experiments *cont.* Category assignment is as follows: very good $(100 - 76\%)$, good $(75 - 51\%)$, bad $(50 - 26\%)$, very bad $(25 - 0\%)$. Percent removal was calculated based on extremes of standard deviations.

Compound Number	**% Removal**	**Category**
21	52	Good[7]
22	90	Very good[7]
23	87	Very good[7]
24	63	Good[7]
25	35	Bad[48]
26	38	Bad[48]
27	100	Very good[7]
28	100	Very good[7]
29	n/A	Inhibitor[20]
30	100	Very good[7]
31	35	Bad[48]
32	46	Bad[7]
33	100	Very good[7]
34	100	Very good[7]
35	21	Very bad[7]
36	7	Very bad[7]
37	100	Very good[7]
38	88	Very good[7]
39	70	Good[7]
40	100	Very good[7]
41	100	Very good[7]
42	58	Good[7]
43	82	Very good[7]
44	14	Very bad[7]
45	95	Very good[7]

Table 2.3: Root mean square deviation (RMSD) of crystal structures with bound ligands with respect to the unbound structure, for the alpha carbons in the active site of NDO expressed in *Pseudomonas* sp. NCIB 9816-4.

PDB Code	**Bound Ligand**	**RMSD (Å)**
1NDO	Unbound	-
1EG9	Indole	0.33
1O7M	O2	0.67
1O7N	Indole	0.51
1O7P	Product	0.32
1O7W	Reduced Unbound	0.54
1O7G	Naphthalene	0.64
1O7H	Unbound	0.34
1UUV	NO + Indole	0.39
1UUW	NO	0.27
2HMK	Phenanthrene	0.57
2HMM	Anthracene	0.57
2HMO	3-nitrotoluene	0.57
4HKV	Benzamide	0.59
4HM0	Indole-3-acetate	0.41
4HM1	1-indanone	0.54
4HM2	Ethylphenylsulfide	0.62
4HM3	Ethylbenzene	0.55
4HM4	Indan	0.57
4HM5	Indene	0.60
4HM6	Phenetole	0.56
4HM7	Styrene	0.58
4HM8	Thioanisole	0.58

Table 2.4: Root mean square deviation (RMSD) of crystal structures with bound ligands with respect to the unbound structure, for the alpha carbons in the active site of BPDO expressed in *Rhodococcus* sp. RHA1

PDB Code	**Bound Ligand**	**RMSD (Å)**
1ULI	Unbound	-
1ULJ	Biphenyk	5.76

Table 2.5: Computational results from our hybrid multi-parameter algorithm. Average ΔG_{AS} and ΔG_{trj} values used to plot Figure 2.8). The percent entrance is defined as the number of frames in which each ligand has *channel continuity* from the entrance (point d from Figure C.3) to $r_\sigma - r_\mu$ (in order to be able to calculate ΔG_{AS} and ΔG_{trj}.

Compound Number	% Entrance	ΔG_{AS}	ΔG_{trj}
1	21	-1.3	9.1
2	72	11.9	6.2
3	11	18.7	1.4
4	21	33.4	4.3
5	2	58.7	24.5
6	13	11.6	16.7
7	39	-43.0	-4.0
8	38	-12.9	7.6
9	61	-12.6	3.7
10	89	-12.7	-40.2
11	98	-12.7	-5.3
12	97	-14.2	-19.4
13	41	-19.0	-24.3
14	94	-20.9	-14.3
15	41	-19.6	-6.2
16	39	-18.6	2.1
17	66	-16.3	-3.6
18	76	-16.0	-1.5
19	21	-10.8	8.2
20	61	-19.9	-31.0

Table 2.6: Computational results from our hybrid multi-parameter algorithm *cont.* Average ΔG_{AS} and ΔG_{trj} values used to plot Figure 2.8). The percent entrance is defined as the number of frames in which each ligand has *channel continuity* from the entrance (point d from Figure C.3) to $r_\sigma - r_\mu$ (in order to be able to calculate ΔG_{AS} and ΔG_{trj}. Compound 44 (triphenylene) did not meet the channel continuity criteria for any frame, therefore values for ΔG_{AS} and ΔG_{trj} were not calculated.

Compound Number	% Entrance	ΔG_{AS}	ΔG_{trj}
21	94	-16.9	-17.4
22	64	-16.5	-6.5
23	72	-20.6	-12.8
24	49	-17.5	-6.5
25	11	30.9	-20.3
26	8	38.9	16.8
27	84	-47.5	-44.7
28	93	-26.8	-40.4
29	43	-60	-12.5
30	85	-32	-18.2
31	19	2.79	-1.1
32	2	62.8	5.3
33	83	-15.7	-1.3
34	37	1.48	7.5
35	9	34.7	17
36	3	19.9	30.4
37	91	-40.4	-13.8
38	97	-27.1	-31.7
39	46	-24.5	-15.5
40	80	-41.3	-18.1
41	78	-49.2	-42.1
42	83	-16.8	-3.7
43	53	-4.95	-2.1
44	0	n/A	n/A
45	51	-19.5	-0.7

Table 2.7: Statistics of water molecules < 10Å from the channel centerline. The averages are based on 100 analyzed frames.

Number of water molecules	10 ± 2
Average distance from centerline (Å)	3.6 ± 0.9

Chapter 3

In silico Identification of Bioremediation Potential: Carbamazepine and Other Recalcitrant Personal Care Products

Adapted with permission from Aukema, K. G., **Escalante, D. E.**, Maltby, M. M., Bera, A. K., Aksan, A., & Wackett, L. P. (2016). In silico identification of bioremediation potential: carbamazepine and other recalcitrant personal care products. *Environmental science & technology*, 51(2), 880-888. doi:10.1021/acs.est.6b04345.

3.1 Chapter Summary

Emerging contaminants are principally personal care products not readily removed by conventional wastewater treatment and, with increasing reliance on water recycling, become disseminated in drinking water supplies. Carbamazepine, a widely used neuroactive pharmaceutical increasingly escapes wastewater treatment and is found in potable water. In this study, a mechanism is proposed by which carbamazepine resists biodegradation and a previously unknown microbial biodegradation was predicted computationally. The prediction identified biphenyl dioxygenase from *Paraburkholderia xenovorans* LB400 as the best candidate enzyme to metabolize carbamazepine. Other recalcitrant personal care products were subjected to prediction by the Pathway Prediction System and tested experimentally with *P. xenovorans* LB400. It was shown to biodegrade structurally diverse compounds. Predictions indicated hydrolase or oxygenase enzymes catalyzed the initial reactions. This study highlights the potential for using the growing body of enzyme-structural and genomic information with computational methods to rapidly identify enzymes and microorganisms that biodegrade emerging contaminants.

3.2 Introduction

Computational approaches to identify enzymes and microorganisms capable of transforming specific compounds are needed to expedite the discovery of bioremediation

potential given millions of biodegradative enzymes available in genomic databases. Current computational tools in the field of biodegradation such as EAWAG-PPS and EPA-BIOWIN are designed to predict the extent of biodegradation and possible metabolites formed via general enzyme classes. In pharmaceutical research, molecular docking and molecular dynamic simulation of enzyme movement are methods routinely used to rationally design small molecule – enzyme interactions. Mammalian P450s have been extensively studied in drug metabolism, including for carbamazepine and other emerging pollutants.[115,116] The current study aims to extend substrate prediction capabilities beyond P450s to other oxygenases known to be important for bioremediation and to catalog substrate predictions in a publically available database, RAPID.

In the present work, computational analysis was used to assess biodegradability of multi-ring compounds like carbamazepine, and a potential explanation for its lack of biodegradability is suggested. Other compounds were tested computationally for their metabolic pathways and the types of enzymes that would react with them. Computational methods led to the identification of a specific biphenyl dioxygenase that rapidly oxidized carbamazepine to non-biologically-active products. *Paraburkholderia xenovorans* LB400, that naturally harbors the reactive biphenyl dioxygenase also contains other oxygenases and hydrolases that biodegrade other recalcitrant personal care products. Some of these compounds have not previously been demonstrated to be biodegradable by a single bacterial strain.

3.3 Methods

3.3.1 Computational Methods

The X-ray structures for the four different enzymes were obtained from the Protein Data Bank (PDB) using the following accession codes: 3EN1 (TDO from *P. putida* F1),[47] 1O7G (NDO from *Pseudomonas sp.* strain 9816-4),[28] 2GBX (BPDO from *S. yanoikuyae* B1),[117] and 2XRX (BPDO from *P. xenovorans* LB400).[118] All of the enzyme structures were prepared using the Schrodinger Protein Preparation Wizard software package.[80] Missing amino acid side chains and hydrogen atoms were added. The partial charges for each of the amino acid atoms were assigned based on the OPLS_2005 force field, except for the iron and 2-his-1-carboxylate facial triad (see §3.3.2). The prepared files were then used for docking using the Glide application in the Schrodinger suite of software.[80] The nonbonded interaction energy (van der Waals and Coulomb) was recorded for all the docking poses. The channel access algorithm is described in Escalante *et al.*[119] Briefly, molecular dynamic simulation of the dioxygenase was carried out for 40 ns. Then using 10% of the static frames or snapshots of the enzyme from the simulation, electrostatic interactions were calculated for the non-bonded interactions of naphthalene and carbamazepine with the tunnel residues as the entrance trajectory into the active site was simulated. After the MD simulations and structure preparations are complete, docking and tunnel simulations can be completed in minutes.

The molecular dynamic (MD) simulation of carbamazepine used the docking pose obtained as the initial input structure. The simulation was run using Desmond to generate an ensemble of energetically accessible structures.[100] The MD system was first relaxed by a series of energy minimizations and short MD simulations, where the temperature of the system was gradually increased from 0 to 300°K, using the default equilibration protocol in Desmond. The production simulation was run for 50 ns at constant temperature and pressure (NPT), where the temperature was maintained at 300°K and pressure at 1 atm.

3.3.2 Partial Charges Calculation

The partial charges for the catalysis motif were calculated using the Gaussian09 program. The motif was first minimized using the B3LYP hybrid functional with the cc-PVDZ basis set. The formal charge of the catalysis motif was defined to be +1 in order to properly model the ferrous oxidation state of iron center. The partial charges were then fitted using the Electrostatic Potential (ESP) scheme included in Guassian09.[120] The same procedure was followed to determine the partial charges of carbamazepine, naphthalene, biphenyl and toluene.

The catalysis motif was able to be successfully minimized to convergence in all three models using the high accuracy basis cc-PVDZ. In order to achieve convergence we froze certain atoms in each of the amino acids, marked by stars in Figure 3.1. For each of the three models the distorted tetrahedral configuration of the atoms

was conserved. The initial parameterization by the force field assigned the partial charge of the ferrous iron to be equal to the formal charge of +2. However as shown in Table 3.1 the mononuclear iron center is actually in a more reduced state with varying degrees for the different models.

For each of the enzymes carbamazepine and the preferred enzyme substrate, naphthalene or biphenyl, were docked. In all of the three models the preferred substrate was able to successfully dock. On the other hand, when carbamazepine was docked the docking score was only favorable for NDO and BPDO.

3.4 Results and Discussion

3.4.1 Computational analysis of Rieske dioxygenases with carbamazepine

Carbamazepine consists of the puckered tricyclic dibenzazepine ring structure with a carboxamide functionality appended to the ring nitrogen atom (Figure 3.2). It is well established to be poorly, if at all, biodegraded in conventional municipal wastewater treatment.[121,10,8] The resistance of carbamazepine to biodegradation has been considered curious since many tricyclic aromatic ring compounds are highly biodegradable, such as anthracene and the structurally-analogous carbazole ring system (Figure 3.2a).[12] The latter compounds are typically biodegraded via initial oxidation by

microbial Rieske oxygenases (Figure 3.2b).[117] Rieske oxygenase genes have been identified in thousands of bacterial genomes and exist on the order of $10^7 - 10^9$ copies/ g sediment.[122,123,124] Moreover, many well-studied Rieske dioxygenases have very broad substrate specificity that we are cataloguing on the RAPID online database to better use natural enzymes for biocatalysis and biodegradation.[125,7]

While Rieske enzymes oxidize hundreds of aromatic compounds, carbamazepine has, to our knowledge, not been tested directly with these enzymes. With thousands of potential Rieske proteins to screen, carbamazepine was investigated with representative dioxygenases computationally first (Figure 3.3). The study was limited to Rieske dioxygenases for which X-ray crystal structures have been solved with and without substrate in the active site. No homology models were used. Of the eight available X-ray structures meeting these criteria, four enzymes with well-characterized substrate specificity were used for computational docking and active site tunnel access. The enzymes and structures are: toluene 2,3-dioxygenase (TDO) from *Pseudomonas putida* F1 (3EN1),[47] naphthalene 1,2-dioxygenase (NDO) from *Pseudomonas sp.* strain 9816-4 (107G),[28] biphenyl 2,3-dioxygenase ($BPDO_{B1}$) from Sphingobium yanoikuyae B1 (2GBX)[117] and biphenyl 2,3-dioxygenase ($BPDO_{LB400}$) from *Paraburkholderia xenovorans* LB400 (2XRX).[118] The two key energy values to identify if a compound fits properly in the active site of the dioxygenase enzymes are the Coulomb and van der Waals (nonbonded) interaction energy. Therefore, if a compound docked in the active site with a positive interaction energy value ($E > 0$kcal/mol), or was unable to be docked by the software due to an excessively high interaction energy, these results would indicate a compound will not be oxidized

by the dioxygenase. By this criterion, TDO was excluded, and $BPDO_{B1}$ appeared unlikely to act on carbamazepine (Figure 3.3a). Rieske dioxygenases have a buried active site, likely providing some control on substrate specificity, and carbamazepine unlike naphthalene was deemed unable to gain entry through the access channel to the NDO using an algorithm developed for Rieske oxygenases (Figure 3.3b). The algorithm calculates the free energy of interaction between the enzyme and the compound as it traverses the channel connecting the solvent region (> 20Å from iron) and the active site pocket (3–10Å from iron). In contrast, $BPDO_{LB400}$ passed all criteria of active site access and productive docking.

The results of molecular docking simulations were used to predict if the presence of the amide functionality might exclude carbamazepine from reacting with many Rieske dioxygenases. The amide group of benzamide, a known inhibitor of NDO, has been proposed to coordinate via the nitrogen through a water molecule to the active site iron atom. Such iron coordination could preclude oxygen binding and thus prevent the catalytic cycle from proceeding. This type of interaction has been observed directly in the 1.65 Å resolution X-ray structure of NDO with benzamide (4HKV).[20] While a non-productive docking orientation with the nitrogen near the iron is favored for the $BPDO_{B1}$ (Figure 3.3c), this is not the case with the $BPDO_{LB400}$. The favored docking pose is consistent with a carbamazepine orientation allowing oxidation of a carbon-carbon double bond (Figure 3.3d). The distance between the iron in $BPDO_{LB400}$ and C3 of carbamazepine is 4.3Å. This distance is in the same range as the equivalent Fe-C distances observed in NDO (1O7G), TDO (3EN1), $BPDO_{B1}$ (2GBX) and $BPDO_{LB400}$ (2XRX) structures with substrate bound. Furthermore, the

equilibrated portion of the molecular dynamic simulation shows that carbamazepine in the active site $BPDO_{LB400}$ was not found to drastically change orientations relative to the initial docking.

3.4.2 Dibenzazepine shown to be more readily oxidized by Rieske dioxygenases

To explore the hypothesis that the carboxamide functionality of carbamazepine is a primary cause of the compound's resistance to biodegradation, parallel experiments were conducted with dibenzazepine, the base ring structure lacking the carboxamide group. Computational analysis suggested that dibenzazepine would dock favorably in the active site of three of the dioxygenases, the exception was toluene dioxygenase (Table 3.2). Unlike carbamazepine, dibenzazepine is predicted to access the active site of NDO, $BPDO_{B1}$ and $BPDO_{LB400}$. Therefore, dibenzazepine was computationally predicted to be a substrate of these enzymes. In wet laboratory experiments, dibenzazepine was accepted as a substrate for three of the strains, the exception being *P. putida* F1 expressing toluene dioxygenase. These data were consistent with the idea that the carboxamide functionality imposed a steric hindrance to being metabolized more readily by microorganisms containing the widely prevalent Rieske dioxygenases. In resting cell assays with 10 ppm dibenzazepine, more than 90% was removed by each of the strains in 24 hours: *Pseudomonas sp.* strain NCIB 9816-4, *Sphingobium yanoikuyae* B1, and *P. xenovorans* LB400.

3.4.3 Biodegradation of personal care products problematic in contemporary wastewater treatment

Carbamazepine is only one of many chemicals escaping current wastewater treatment facilities. In light of the unique ability of *P. xenovorans* LB400 to metabolize carbamazepine rapidly, its unusually large genome (9.73 Mb), its unusually high number of predicted oxygenases and degradative enzymes (Table 3.3 and 3.4) and its broad catabolic activities with PCBs, dioxins and terpenoids,[126,127,128] other chemicals of emerging concern were examined here. These included other poorly-degradable pharmaceuticals, endocrine disrupting alkyl phthalates, a sunscreen agent, fragrance compounds and detergents, many of which are increasingly appearing in municipal water supplies. Each compound was incubated at a concentration of 10 ppm in resting whole cell assays for 24 hours with *P. xenovorans* LB400 and *E. coli* control are shown in Figure 3.4a. The full list of compounds tested and the GC/FID and HPLC data are provided in Tables 3.5 and 3.6. Chromatogram peak identity was confirmed by comparison of retention time to authentic standards. For compounds volatile enough to be analyzed by GC, further confirmation of peak identity was provided by the mass spectrum.

Because P. xenovorans LB400 is known to express a large number of biodegradative enzymes (Tables 3.3 and 3.4), it is unlikely that all of these compounds are substrates of BPDO. While a full analysis of the biotransformation route of each compound is beyond the scope of this study, each compound was computationally

analyzed to predict the likely initiation enzyme(s). The EAWAG-Pathway Prediction System (PPS) is a prediction approach that computationally predicts metabolic pathways using a rule-based system generated by expert knowledge and gene frequencies.[129,130] The PPS shows the most likely initiating reaction types and is complementary to the RAPID prediction algorithm being developed since the PPS is not capable of predicting a specific enzyme. The compounds degraded were clustered by percent removed and by the enzyme-types initiating the metabolism of the compounds as predicted by the PPS. (Figure 3.4). Two observations can be made from the clustering. First, clustering of the compounds by percent removal alone clearly shows that chemical structure rather than contaminant source determines the extent of degradation. Second, clustering by the two parameters shows that those compounds transformed >50% are computationally predicted to undergo biotransformation via esterase or amidase enzymes. Likewise, the more recalcitrant compounds, those degraded < 50% in 24 hours, are predicted to be biotransformed via mono- or dioxygenases. In total, we tested 22 emerging pollutants, 10 of which were not significantly degraded (Figure 3.5).

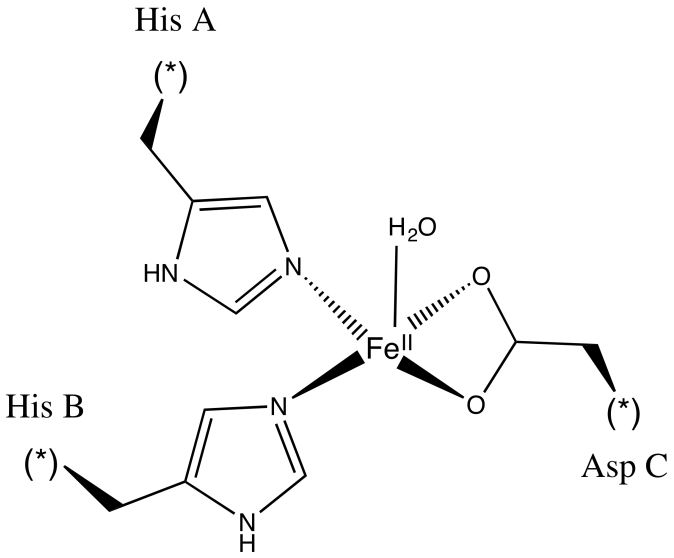

Figure 3.1: General structure of the 2-his-1-carboxylate facial triad along with iron center found in Rieske non-hemed dioxygenases.

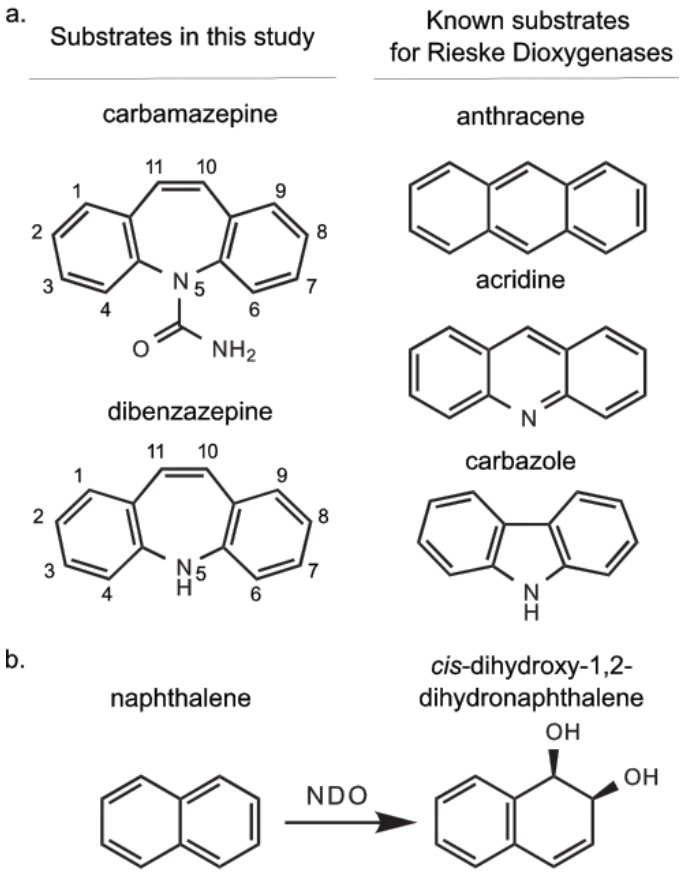

Figure 3.2: (a) Carbamazepine and common Rieske dioxygenase substrates. The numbering system for dibenzazepine ring atoms follow the accepted numbering as indicated in the figure. (b) A typical reaction of a polyaromatic hydrocarbon with a Rieske dioxygenase, naphthalene dioxygenase (NDO).

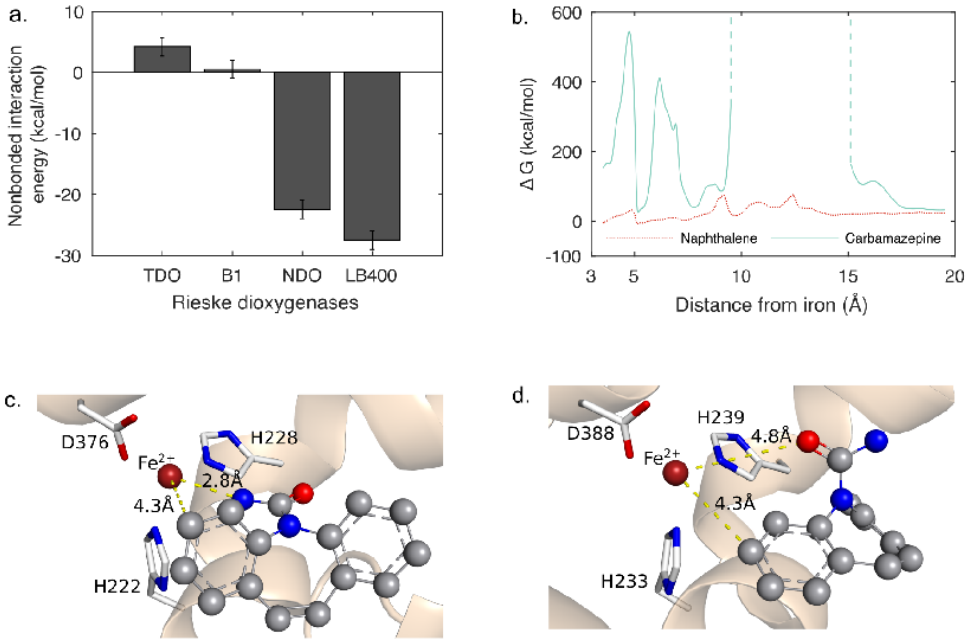

Figure 3.3: Computational analysis of the interaction between Rieske dioxygenases and carbamazepine to predict enzymatic reactivity. (a) Nonbonded interaction energy between carbamazepine and active site of four Rieske dioxygenases. (b) Free energy of interaction between carbamazepine and the NDO channel leading into active site pocket as a function of compound position in the channel. Distances are measured from iron at the distal end of the active site. Naphthalene, the natural NDO substrate, is shown for comparison. (c,d) Docking position of carbamazepine in $BPDO_{B1}$ and $BPDO_{LB400}$, respectively. Only $BPDO_{LB400}$ passed all computational requirements for reacting with carbamazepine.

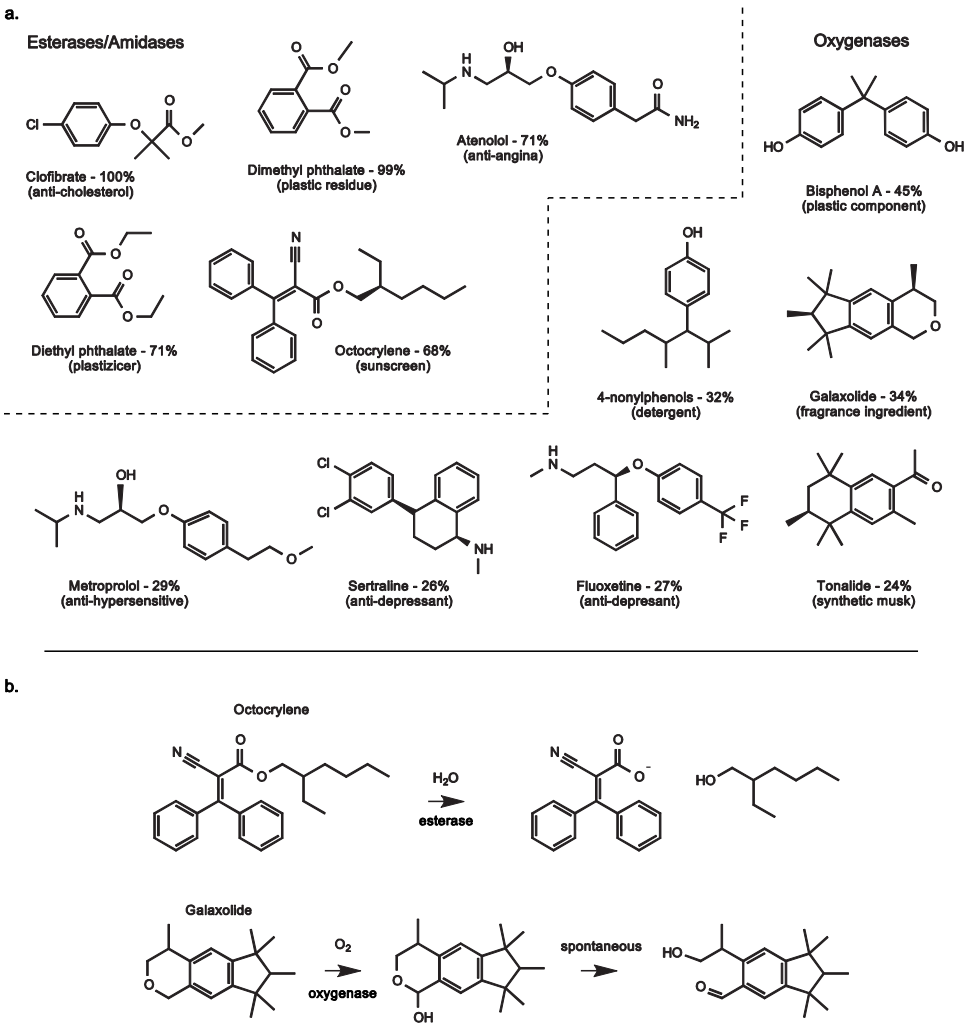

Figure 3.4: (a) Biotransformation of emerging pollutants by *P. xenovorans* LB400 after 24hr. Percent removal of each compound relative to an *E. coli* control is indicated after the name of the compound. The commercial application of the compound is shown in parantheses. Pathway Prediction System (PPS) models were used to determine the likely initiating metabolism, denoted by esterases/amidase to the left of the dotted lines and by oxygenases to the right. (b) Example esterase- and oxygenase-mediated degradation pathways predicted by the PPS for two representative emerging pollutants.

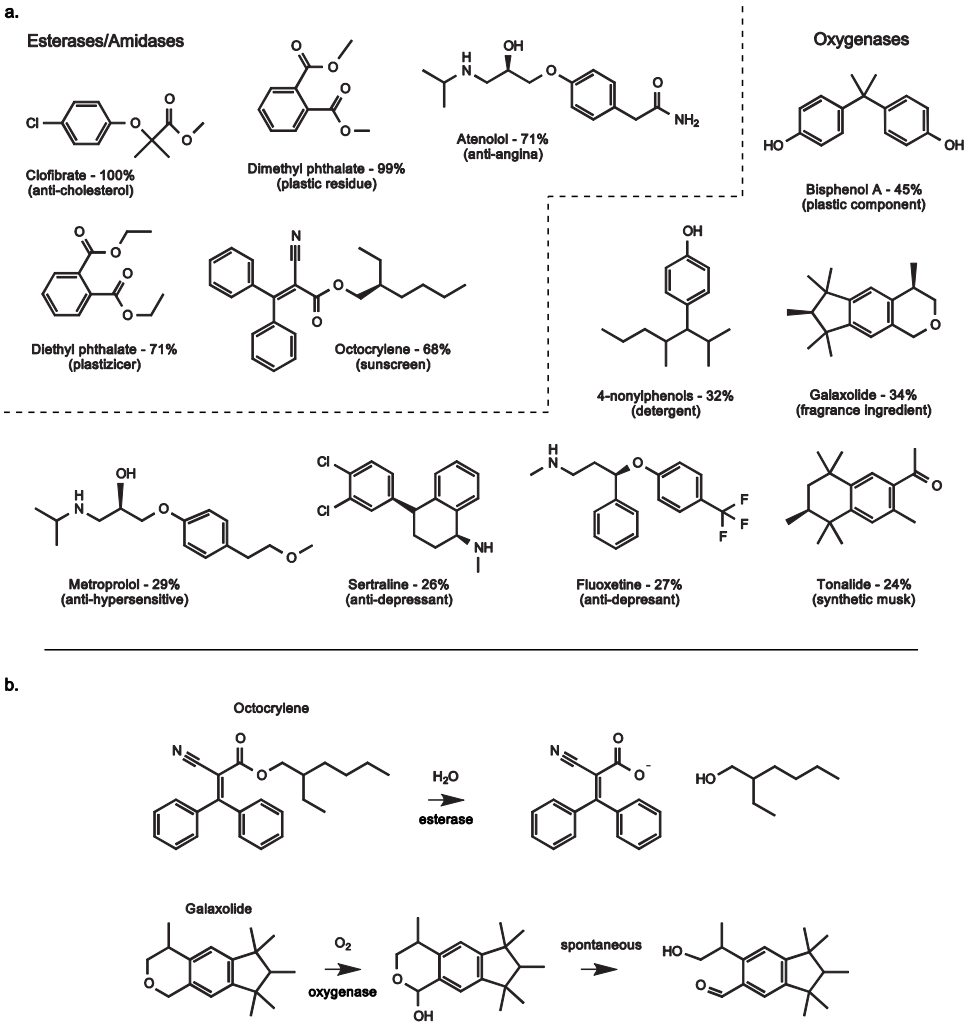

Figure 3.5: Emerging pollutants not significantly degraded by *P. xenovorans* in a 24 hr resting cell assay.

Table 3.1: Partial Charges of 2-his-1-carboxylate facial triad in four different non-heme Rieske dioxygenases

	Residue Number			Partial Charges (q)	
	His A	His B	Asp C	Fe(II)	H-O-H
$NDO_{9816\text{-}4}$	208	213	362	1.304	0.433 – (−0.813) – 0.411
$BPDO_{B1}$	207	212	360	1.339	0.424 – (−0.814) – 0.447
$BPDO_{LB400}$	233	239	388	1.544	0.422 – (−0.840) – 0.432
TDO	222	228	376	1.028	0.417 – (−0.752) – 0.414

Table 3.2: Nonbonded energy of dibenzazepine in the four Rieske dioxygenases studied

	Energy (kcal/mol)		
Rieske Dioxygenase	Coulomb	van der Waals	Total (n.b)
TDO	-1.5	1.9	0.4
BPDOB1	-0.9	-28.8	-29.7
BPDOLB400	-0.1	-24.4	-24.5
NDO	-2.3	-32.2	-34.5

Table 3.3: Compounds biotransformed by *P. burkholderia*

Compounds	References
biphenyl	[a] 10.1128/AEM.01129-06
polychlorinated biphenyls	10.1128/AEM.01129-06
abeitic acid	10.1128/JB.00179-07
dehydroabeitic acid	10.1128/JB.00179-07
palustric acid	10.1128/JB.00179-07
7-oxo-dehydroabeitic acid	10.1128/JB.00179-07
gentisate	10.1371/journal.pone.0056038 10.1371/journal.pone.0017583
protocatechuate	10.1371/journal.pone.0056038 10.1371/journal.pone.0017583
chloroacetaldehyde	10.1128/AEM.01129-06
taurine	10.1042/BJ20021455
p-cymene	Montes-Matias thesis with G. Zylstra, 2008 Rutgers University
formaldehyde	[b]
phenylacetic acid	10.1007/s00203-011-0705-x 10.1371/journal.pone.0017583
benzoate	[a]
2-aminophenol	10.1371/journal.pone.0075746

[a]– 10.1128/AEM.70.8.4961-4970.2004 [b]– 10.1128/JB.186.7.2173_2178.2004

Table 3.4: Compounds predicted to be biotransformed by *P. burkholderia* based on genome

benzonitrile via catechol
benzamide via catechol
mandelate via catechol
benzaldehyde via catechol
anthranilate via catechol
salicylate via catechol
vanilline via protocatechuate
4-hydroxybenzoate via protocatechuate
phtalate via protocatechuate
terephthalate via protocatechuate
4-carboxydiphenyl ether via protocatechuate
3-chlorocatechol via protocatechuate

Reference: 10.1073/pnas.0606924103

Table 3.5: GC/FID peak area of compounds following 24 hour resting cell assay with *P. xenovorans* LB400 or E. coli.

	GC/FID peak area x 10^6		
chemical name	*P. xenovorans* LB400	*E. coli* DH5α	% removed
clofibrate	not detected	3.9 ± 0.6	100
dimethyl phthalate	0.14 ± 0.02	12.5 ± 0.4	99
diethyl phthalate	0.37 ± 0.03	1.3 ± 0.1	71
octocrylene	0.15 ± 0.03	0.47 ± 0.12	68
bisphenol A	1.1 ± 0.4	1.9 ± 0.1	45
galaxolide	0.9 ± 0.3	1.4 ± 0.3	34
4-nonylphenols	1.9 ± 0.3	2.7 ± 0.2	32
fluoxetine	1.4 ± 0.2	1.9 ± 0.2	27
sertraline	0.81 ± 0.01	1.1 ± 0.1	26
tonalid	3.3 ± 0.7	4.4 ± 0.4	24
tiabendazole	0.8 ± 0.1	0.8 ± 0.1	<15
DEET	5.2 ± 0.3	5.4 ± 0.4	<15
dioxane	0.48 ± 0.02	0.45 ± 0.01	<15
primidone	0.72 ± 0.04	0.69 ± 0.02	<15
cotinine	0.25 ± 0.02	0.24 ± 0.01	<15

Table 3.6: HPLC peak area of compounds following 24 hour resting cell assay with *P. xenovorans* LB400 or E. coli.

	HPLC peak area x 10^3		
chemical name	*P. xenovorans* LB400	*E. coli* DH5α	% removed
atenolol	1 ± 0.1	3.4 ± 0.1	71%
metoprolol	1.2 ± 0.1	1.7 ± 0.1	29%
sulfamethoxazole	1.8 ± 0.1	1.9 ± 0.1	<15%
gemfibrozil	2.8 ± 0.1	3.1 ± 0.1	<15%
lamotrigine	4.6 ± 0.2	4.6 ± 0.2	<15%
sulfathiazole	2.1 ± 0.1	2.1 ± 0.1	<15%
trimethoprim	2.1 ± 0.1	1.9 ± 0.1	<15%

Chapter 4

Role of Water Hydrogen Bonding on Transport of Small Molecules Inside Hydrophobic Channels

Adapted with permission from **Escalante, D. E.**, & Aksan, A. (2019). Role of water hydrogen bonding on transport of small molecules inside hydrophobic channels. *The Journal of Physical Chemistry B.* doi:10.1021/acs.jpcb.9b03060.

4.1 Chapter Summary

We present a systematic analysis of water networking inside smooth hyperboloid hydrophobic structures (cylindrical, barrel, and hourglass shapes) in order to elucidate the role water hydrogen bonding on transport of small hydrophobic molecules (ligands). Through a series of molecular dynamic simulations, we establish that a hydrogen bonded network forming along the centerline results in a water exclusion zone adjacent to the walls. The size of the exclusion zone is a function of geometry and non-bonded interaction strength; defining the effective hydrophobicity of the structure. Exclusion of water molecules from this zone results in a lower apparent viscosity, leading to acceleration of ligand transport up to seven times that measured in the bulk. Access of the ligands into and out of the hydrophobic structures was found to be controlled by single water molecules that cap regions of small diameter. This capping mechanism provides physical insights into the behavior and role of water at the bottleneck regions of real hydrophobic biological channels. The set of geometries are then used to develop a model that can predict transport of ligands along nanochannels of broad-substrate specificity enzymes.[131]

4.2 Introduction

Many enzymes have channels that control the bidirectional transport of small organic molecules (ligands) from the solvent-exposed surface into their buried catalytic cavity

(active site).[24] Experimental and computational results have shown that enzymes that have channels can exhibit broad-substrate specificity (BSS)[7,132,26,133,19,13,27,94,134,135,16,136,137,17,56] they can perform the same reaction with hundreds, even thousands, of different ligands *if* the ligand reaches the active site. A ligand can reach the active site only if: i) nonbonded interactions with the channel wall do not induce high-energy barriers;[119,138] ii) a gate or bottleneck does not physically block its transport;[139] and iii) it can displace water molecules within the channel.[119] Two examples of BSS enzymes are: i) cytochrome P450s, known to have a central role in drug metabolism and detoxification of xenobiotics;[13,94,140,137,17,141] and ii) Rieske non-heme iron oxygenases (known to start biodegradation of many recalcitrant pollutants), which are used to synthesize industrially and medically relevant chiral chemicals[27,7,26,19,27,142,135] An experimental exploration of new substrates for BSS enzymes can be costly and laborious.[18] Therefore, computational tools that can expedite the discovery of new substrates are needed.

Computational tools, based on molecular dynamic (MD) simulations, have been successful at modeling the transport of ligands along BSS enzyme channels.[18,119] For instance, our group developed an algorithm that analyzes the movement of ligands through the channels by calculating the non-bonded pairwise potentials for all atoms in any given ligand/enzyme pair.[119] Steered molecular dynamics (SMD), and random acceleration molecular dynamics (RAMD) were also used to calculate the energetic profile of a ligand that is being pulled by an external force along a channel[14,143,144,145,16,146,54,108] An advantage of SMD and RAMD simulations is that they provide information about the ligand transport pathway without the need for *a priori*

knowledge about the channel.[18] However, all MD-based methods require calculation of interaction energies between every possible pairwise combination. This increase the computational cost substantially to $10^3 - 10^5$ seconds per ligand per enzyme[131] In an attempt to reduce the computational time, some simulations have used implicit solvent models[147] while others have ignored the presence water molecules inside channels.[119] These approaches therefore, do not account for the effect of water-water hydrogen bonding on ligand transport, which we show here to be a very important factor. Hence, all-atom methods, despite their demonstrated success at a small scale, are not suitable for high-throughput high accuracy screening and prediction of substrates.

Coarse-grained (CG) models are an alternative to all-atom methods since they can reduce the computational burden while still including explicit solvent molecules. Certain CG models developed for enzyme channel analysis use cylindrical carbon nanotubes (CNTs) to simulate the behavior of water in the presence of external forces (e.g. pressure or electric field gradients). See Chakraborty et al[148] for a review of these models. Others include transport of ions along cylindrical CNTs.[149,150,148,151,152,153,154] Therefore, CNTs have served as successful prototypes for modeling transport of small molecules along certain transmembrane channels, such as ion channels, and aquaporins.[155] However, the current CG models are not adequate to study transport of ligands along the channels of BSS enzymes. The majority of BSS enzyme are globular[119] their substrates are uncharged hydrophobic molecules,[27,7] and the channels seldom have a cylindrical geometry[24] Therefore, non-cylindrical hydrophobic CG models are required to properly study the transport of ligands in BSS enzymes.

Knowledge of the behavior of water surrounding the hydrophobic ligands inside non-cylindrical channels is limited.[156] Some general insights can be derived from the extensive research conducted on the behavior of water surrounding hydrophobic bodies in solution.[157,158,159,160,161,162,163,164,165,166] For instance, solvating a single methane molecule (i.e. small hydrophobic ligand) generates a spike in the density of water near its vicinity.[160,161] This spike is caused by a crowding effect that allows formation of a complete solvation shell around the methane molecule. The solvation shell maximizes the number of hydrogen bonds (HB) for each water molecule, compensating for the loss of enthalpy due to the solvation process.[162,163,164] On the other hand, water behaves differently when large hydrophobic bodies are solvated for example, when hydrophobic flat plates at the same length scale as enzyme channels are placed in water, the water molecules surrounding the plates are unable to reorganize in such a way that the maximum number of HBs are attained.[165,166] In this case, on average, each water molecule could only hydrogen bond with two of its four-possible nearest-neighbors, causing an enthalpic penalty for solvation, since at least one hydrogen atom has to point towards the hydrophobic surface.[165,166] The differences in solvation enthalpy between small and large hydrophobic bodies/surfaces suggest that the geometry of a channel is an important factor in determining the behavior of water, and therefore the ligand solvated in it.

In the present study, we examined the behavior of water and hydrophobic ligands inside non-cylindrical channels through the use of MD simulations. To systematically study this phenomenon, we defined a set of building blocks (BB) described by their: i) geometrical shape; and ii) level of wall hydrophobicity. Three different

geometries were chosen to simulate cylindrical and non-cylindrical channels. And six hydrophobicity levels were chosen to cover the full range identified by the developers of MOLE 2.0.[23] This resulted in a total of 18 different building blocks that can be used interchangeably to model, both the geometry and hydrophobicity of, BSS enzymes channels as described below.

The chosen BB geometries were: a cylinder, a barrel and an hourglass (Figure 5.2). Although the behavior of water inside cylindrical channels is well understood,[148] we included the analysis of cylinders in our study to be able to compare our results to the literature and have a full set of BBs simulated under same conditions. The two non-cylindrical geometries were chosen based on crystallographic descriptions of BSS enzyme channels. It has been demonstrated that the channel leading to the entrance of the active site in BSS enzymes 'is similar to an inverted funnel, with a small aperture leading to a large vestibule,[21] i.e., an hourglass BB followed by a barrel BB. Furthermore, MD simulations have shown that many BSS enzyme channels are mostly rigid; except for very narrow regions of localized flexibility due to the movement of single amino acids acting as hinge-gates.[7,119,139] Therefore, we propose that the architecture of a BSS enzyme channel can be digitized as series of BBs that are sequentially connected to one another. The digitization technique we developed here has been successfully applied to the channel of naphthalene 1,2-dioxygense, a BSS enzyme in a very recent parallel publication.[131]

4.3 Methods

4.3.1 The Building Block Simulation Setup

We used three different geometries (Figure 5.2); a cylinder, a convex hyperboloid (barrel), and a concave hyperboloid (hourglass). The walls of the BBs were constructed using neutral particles (pseudoatoms) organized in a hexagonal closed packing (hcp) grid aligned with the z-axis at a lattice constant of 1.2Å. All BBs are of the same length $L = 20$Å, their entrance and exit radii are denoted as $r(z = \pm 10\text{Å}) = r_i$, and the radius at the middle of the BB is denoted $r(z = 0) = r_o$ (see Table 4.1 for numerical values for each BB). The surface of a hyperboloid is given in Cartesian coordinates by Equation 4.1:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1 \tag{4.1}$$

where $a^2 = b^2 = r_o^2$ for a radially symmetrical channel as it is in our case. Equivalently, equation 4.1 can be represented in cylindrical coordinates by Equation 4.2:

$$\frac{r^2}{r_o^2} - \frac{z^2}{c^2} = 1 \tag{4.2}$$

and the geometry constant c is given by Equation 4.3:

$$c = \frac{r_0 \cdot L}{\sqrt{(r_i + r_o)(r_i - r_o)}} \tag{4.3}$$

In order to obtain building blocks with wall thickness of two atoms we deleted all the atoms where $r > 5$ and $r < 7$.

The ligand was modeled using a sphere of radius $r_L = 3$Å and ten ligands were initially placed in the upper bulk reservoir ($z > 10$Å). In order to guide the ligands into the BB, two plates perpendicular to the z-axis were placed at the entrance and the exit regions as shown by the blue atoms in Figure 4.2a-c. Periodic boundary conditions were imposed in three axes. To avoid molecules from recirculating back to the upper reservoir, a third impermeable plate was positioned 15Å above the entrance of the BB (at $z = 25$Å) as shown by red atoms in Figure 4.2c.

We parameterized the BB wall as a hydrophobic surface by changing the depth, ε_C, of the Lennard-Jones potential well. The range of Lennard-Jones potential well values were chosen to match the hydrophobicity levels of the enzyme channels as calculated by the channel identification program MOLE 2.0.[23] Six different values were chosen and normalized against the Lennard-Jones potential well value of water, $\varepsilon_W = 0.1554$ kcal/mol,[43] resulting in building blocks with potential well ratios (PWR) of $\varepsilon_C/\varepsilon_W = 0.5, 0.6, 0.7, 0.8, 0.9, 1.0$. Since our model is developed to predict transport of small hydrophobic compounds, hydrophobicity of the ligand was kept constant in all simulations at $\varepsilon_L = 0.07$ kcal/mol, representing a general sp^2 hybridized CH atom similar to a methane molecule with a van der Waals radius of $\sigma_L = 1.4$Å.[45] Each of the building blocks was solvated using single point charge (SPC) water molecules,[167] resulting in a simulation box centered at the origin with dimensions of 30Å × 30Å × 100Å. The force field values for the water molecules are given in Table 4.1. To

ensure equilibration of the water molecules within and outside of the BB, a grand canonical Monte Carlo method was used to sample the number of water molecules in the building block and their positions before any MD simulations were run. The Monte Carlo method was performed using the solvate-pocket utility in Desmond.[168]

After ensuring that water molecules inside all BBs were equilibrated, MD simulations were performed using the Desmond Molecular Dynamics Simulation package[168] under the *NPT* ensemble. The temperature was kept constant at 298K using the Nose-Hoover chain thermostat with a relaxation time of 1ps. The pressure was kept constant at 1atm using the isotropically-coupled Martyna-Tobias-Klein barostat with a relaxation time of 2 ps. The RESPA integrator algorithm was used with time steps of 2, 2, and 4 fs for bonded, near, and far interactions, respectively. Short range Coulombic interactions were treated using a cutoff of 9Å and long-range interactions were treated using the particle mesh Ewald with a tolerance of $1 \times 10^{-}9$. All pseudoatoms forming the BBs walls and the boundary plates were kept in place by applying a harmonic constraint of 100 kJ/mol $\cdot$ Å^2. Initial velocities were assigned from a Boltzmann distribution (298 K) followed by 1 ns of equilibration dynamics with velocities being reassigned every 0.1 ps, the production simulation was then ran for 20 ns with no further velocity reassignment and recording trajectory snapshots at 1 ps intervals, for a total of 20000 frames per building block geometry per PWR. The first 5 ns of the production simulation were taken as equilibration and the remaining 15 ns were used for analysis, except for the barrel and hourglass BBs at PWR = 0.5 and 0.6, where only the last 10 ns were used for analysis.

4.3.2 Evaluation of thermodynamic properties

The following thermodynamic properties were evaluated for the water and ligand molecules inside the BBs: intermolecular interaction potential (U), excess entropy (S_x), Gibbs free energy (G), diffusion coefficient (D), average residence time ($\bar{t}_{\text{res}}$), average adsorption time ($\bar{t}_{\text{ads}}$). Water molecules were determined to be inside the BB if the oxygen atom met the following criterion:

$$\frac{r^2}{r_o^2} - \frac{z^2}{c^2} < 1 \tag{4.4}$$

in the range $-10 \leq z \leq 10$, where c is the geometrical constant given in . The same criterion was used to determine if the ligand molecules were inside the BB. This allowed us to define a set of water, and ligand molecules for each time step ($\mathbb{W}$, and $\mathbb{L}$ respectively), such that n is the frame number being analyzed. A fixed set $\mathbb{C}$ was defined for all of the atoms making up the BB; note that the subscript n has been dropped since their positions were constrained during the simulations. In order to track the position of each atom in $\mathbb{W}$ and $\mathbb{L}$ we constructed a three-dimensional binning matrix ($\mathbb{B}$) for the region $-5 \leq x, y \leq 5$ and $-10 \leq z \leq 10$ where each bin (b), such that $b \in \mathbb{B}$, covered a distance of 0.01Å in each Cartesian direction, i.e., the matrix has dimensions of $1000 \times 1000 \times 2000$.

The interaction potential energy for each bin (b) at a given frame (n) inside

the building block was defined as follows:

$$U(b,n) = U_{\mathrm{WW}} + U_{\mathrm{WL}} + U_{\mathrm{WC}} + U_{\mathrm{LW}} + U_{\mathrm{LL}} + U_{\mathrm{LC}} \tag{4.5}$$

where, where the subscripts W, L, and C correspond to the set of water, ligand, and building block atoms, respectively. The total intermolecular pairwise addition for each term in Equation 4.5 can be generalized in the following way:

$$U_{ij}(b,n) = \sum_{i}^{\beth \in b} \sum_{j}^{\forall \mathbb{B}} \frac{q_i q_j}{r^2 e} + 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{4.6}$$

where $\beth$ represents any of the sets ($\mathbb{W}_n$ or $\mathbb{L}_n$) such that the atoms in this set are found inside bin b, and j can represent any of the three subsets ($\mathbb{W}_n$ or $\mathbb{L}_n$ or $\mathbb{C}$) for all bins b during the time frame n. We then used the following equation to calculate the ensemble average of the potential energy for each bin b as follows:

$$\langle U_{ij}(b) \rangle = \frac{\sum_{n=1}^{N} U_{ij}(b,n) e^{-U_{ij}(b,n)\beta}}{\sum_{n=1}^{N} e^{-U_{ij}(b,n)\beta}} \tag{4.7}$$

where $\beta = (k_{\mathrm{B}}T)^{-1}$, and k_{B} is the Boltzmann constant and $T = 298$ K. Equation 4.7 is used to determine the potential energy at any given point inside each BB and plot the total ensemble average interaction potential energy along the radial axis of the BBs.

In order to calculate the excess entropy of a single water molecule ($S_{i,x}$) inside each BB we used the permutation reduction method.[169,170,171] In this procedure a linear assignment approach is used to assign a reference initial position to each

water molecule, causing each permuted water molecule to move only in a localized region. We calculated the 3×3 mass-weighted covariance matrix of translational fluctuation along the Cartesian coordinates of each permuted water molecule for all n frames. For each permuted water molecule (j) inside the building block, we diagonalized the covariance matrix resulting in three eigenvalues ($\lambda_i \mid i = 1 - 3$) and their corresponding quasiharmonic frequencies $\omega_i = \sqrt{k_{\mathrm{B}}T/\lambda_i}$. We then calculated the two-particle translational entropy using the frequency values for each water molecule j as follows:[171,172]

$$S^{(2)}_{\mathrm{trans},j} = k_{\mathrm{B}} \sum_{i=1}^{3} \frac{\hbar\omega_i/k_{\mathrm{B}}T}{e^{\hbar\omega_i/k_{\mathrm{B}}T} - 1} - \ln\left(1 - e^{-\hbar\omega_i/k_{\mathrm{B}}T}\right) \tag{4.8}$$

where $\hbar = h/2\pi$ and h is the Plank's constant. The two-particle rotational entropy was calculated using the angular distribution of each permuted water molecule, j, according to the following expression:[171]

$$S^{(2)}_{\mathrm{rot},j} = -k_{\mathrm{B}} \int p(\theta,\chi) w \ln\left(p(\theta,\chi) w\right) \sin\theta \, d\theta \, d\chi \tag{4.9}$$

where $p(\theta,\chi)$ is the angular distribution of each permuted water molecule over all n time frames, w is a normalization factor such that $\sum p(\theta,\chi) = 1$, and θ and χ are two angles describing the rotational orientation of each water molecule (See Section 4.3.3 for more details). Lazaridis and Karplus[173] show that the entropy due to two-particle interactions is simply the sum of translational and rotational components: $S^{(2)} = S^{(2)}_{\mathrm{trans}} + S^{(2)}_{\mathrm{rot}}$. This two-particle term has been shown to account for up to 95% of the total excess entropy of a system.[173] Therefore, we estimated the excess entropy

of each single water molecule j inside the BBs as follows:

$$S_{x,j} \approx S_j^{(2)} = S_{\text{trans}}^{(2)} + S_{\text{rot}}^{(2)} \tag{4.10}$$

We then calculated the total excess entropy (S_x) inside the BB as the ensemble average over all j water molecules, using:

$$S_x = \frac{\sum_j^{\mathbb{W}_{\text{ref}}} S_{x,j} e^{-S_{x,j}/k_{\text{B}}}}{\sum_j^{\mathbb{W}_{\text{ref}}} e^{-S_{x,j}/k_{\text{B}}}} \tag{4.11}$$

where $\mathbb{W}_{\text{ref}}$ is the set of water molecules inside each building block used as a reference for the permutation reduction procedure. The ensemble average Gibbs free energy ($\langle G \rangle$) can be calculated by substituting Equations 4.7 and 4.11 into the definition of Gibbs free energy:

$$\langle G \rangle =] \langle U \rangle - TS_x \tag{4.12}$$

In order to calculate the diffusion coefficient of water (D_W) and ligands (D_L), we tracked the position of each particle residing inside the building block. Each residence event was defined as the series of consecutive time frames during which the particle meets the criteria imposed by Equation 4.4 for the range $-10 \leq z \leq 10$. The mean square displacement (MSD) for the type of particle (water or ligand) was calculated using Equation 4.13:

$$\langle \mathbf{x}^2(t) \rangle = \frac{1}{N_k} \left| (\mathbf{x}_k(t) - \mathbf{x}_{0,k})^2 \right| \tag{4.13}$$

where the vector $\mathbf{x}_k(t)$ represents the Cartesian coordinates of the particle at time t for each residence event k, $\mathbf{x}_{0,k}$ represents the Cartesian coordinates of the particle at time 0 for each residence event k, and N_k is the total number of residence events. The general formula to compute the diffusion coefficient is given by:

$$\langle \mathbf{x}^2(t) \rangle = 6Dt^{\alpha} \tag{4.14}$$

where D is the diffusion coefficient, and α defines the type of diffusion mechanism. This parameter can take the values of $\alpha = 1$ for Fickian diffusion, $0 < \alpha < 1$ for confined diffusion, or $\alpha > 1$ for ballistic diffusion.[174]

The average residence time ($\bar{t}_{\text{res}}$) of the ligands was calculated as the average duration of all ligand residence events inside the BB. In addition, we defined that a ligand was adsorbed onto the BB wall if the condition imposed by Equation 12 and $-10 \leq z \leq 10$ were met:

$$\left(x_i^2(t) + y_i^2(t)\right)^2 > r_o \left(1 + \frac{z_i^2(t)}{c^2}\right) - 1.2\sigma_L \tag{4.15}$$

where x_i, y_i and z_i are the Cartessian coordinates for each of the ten ligands and σ_L is the van der Waals radius of the ligands. We then determined all consecutive frames for which the adsorption conditions were met and averaged them to calculate the average adsorption time ($\bar{t}_{\text{ads}}$). Finally, the volume fraction (V^*) was defined as the average volume occupied by the water molecules ($\bar{V}_{\text{W}}$) per total volume of the

building block (V_{B}), as defined by Equations 4.16 and 4.17:

$$\bar{V}_{\mathrm{W}} = \frac{\sum V\left(\mathbb{W}_n\right)}{N} \tag{4.16}$$

where $V\left(\mathbb{W}_n\right)$ is the volume calculated by plotting the convex hull of all atoms belonging to the set $\mathbb{W}_n$ and N is the total number of frames. And the total volume of the building block can be calculated using:

$$V_{\mathrm{B}} = \frac{2\pi}{3} r_o^2 L \left(3 - \frac{L^2}{c^2}\right) \tag{4.17}$$

4.3.3 Calculation of Water Angular Distribution Angles

The angles θ and χ are used to describe the rotational orientation of each water molecule. The definition of each angle is as follow:

$$\theta = \cos^{-1}\left(\frac{r_{\mathrm{lw}} \cdot \mu}{|r_{\mathrm{lw}} \cdot \mu|}\right) \tag{4.18}$$

$$\chi = \cos^{-1}\left(\frac{h \cdot n}{|h \cdot n|}\right) \tag{4.19}$$

Vector r_{lw} is the distance from the oxygen ($\overline{x}_{\mathrm{oxygen}}$) atom to the origin:

$$r_{\mathrm{lw}} = |\overline{x}_{\mathrm{oxygen}}| \tag{4.20}$$

Vector h is the direction of from hydrogen 1 ($\overline{x}_{\text{hydrogen-1}}$) to ($\overline{x}_{\text{hydrogen-2}}$):

$$h = \overline{x}_{\text{hydrogen-2}} - \overline{x}_{\text{hydrogen-1}} \tag{4.21}$$

Vector μ is the direction of the water molecule dipole moment:

$$\mu = (-0.82 \times \overline{x}_{\text{oxygen}}) + (0.41 \times \overline{x}_{\text{hydrogen-1}}) + (0.41 \times \overline{x}_{\text{hydrogen-2}}) \tag{4.22}$$

Vector n is perpendicular to r_{lw} and μ:

$$n = r_{\text{lw}} \times \mu \tag{4.23}$$

4.4 Results

To understand the behavior of water molecules when confined within the BBs, and to establish their effect on the transport of ligands, we carried out two sets of simulations: the first set included BBs of different geometry flooded with water molecules (the "-" set), while the second set included ten ligands in addition (the "+" set), initially placed in the upper reservoir (Figure 5.2a). Each set of simulations explored six different PWRs ranging between $\varepsilon_{\text{C}}/\varepsilon_{\text{W}} = 0.5 - 1.0$ at 0.1 intervals for each BB geometry (i.e., a total of 18 different conditions). Note that as PWR increases, the attraction force between the BB wall and the water molecule increases, making it a less hydrophobic interaction. In order to validate our simulations with the SPC water

model, we calculated the density and viscosity of the bulk water in the lower water reservoir in all (–) simulations as $\rho = 0.0334\text{Å}^{-3}$ and $\eta = 0.471 \times 10^{-3}$ kg m^{-1} s^{-1}. Both of these values are within 5% of the values reported in the literature.[175,176]

4.4.1 Water and Ligand Density Profiles

Figure 4.3a shows the results of (-) simulations for the axial density distribution of water inside three different BBs at varying PWRs. The distributions are normalized with respect to bulk water. In cylindrical BBs at all PWR values, we observed formation of peaks with a period of $\sim$ 2.4Å (Figure 4.3a.1). This is lower than the 2.8Å value measured for bulk water[175] but is close to the $\sim$ 2.5Å value observed inside $(6, 6)$ carbon nanotubes (CNT) of comparable diameter[149] The lower peak period in the BB compared to bulk indicate an increase in the ordering of water molecules in the cylindrical BB. As PWR decreased, local density decreased without any drastic change in peak shape and periodicity (Figure 4.3a.1). A similar behavior of the water molecules inside the barrel BB were observed at the three highest PWRs (Figure 4.3a.2), albeit with a larger period of $\sim$ 2.9Å. On the other hand, we observed formation of a low-density region between $z = \pm 7$Å for the two lowest PWR values (i.e. the two most hydrophobic cases) in the barrel BB in (–) simulations. The low-density region is characterized by the absence of any discernible peaks. Figure 4.3a.3 shows that for the hourglass (-–) BB simulations at the five highest PWR values the peak period was $\sim$ 1.7Å, lower than the period observed in the cylindrical (–) BB simulations at the same PWR. This indicates that water molecules inside the

hourglass (-) BB have the highest degree of ordering out of all three BB geometries.

Next, we explored the local density profile of water inside the three BB geometries at varying PWRs for the (+) simulations in the presence of ligands (Figure 4.3b), as well as the ligand density profiles inside all BBs (Figure 4.3c). Note that all profiles are normalized with respect to the density of bulk water. We observed that the distribution of water molecules inside the cylindrical (+) BB and barrel (+) BB at the highest PWRs was not affected by the presence of ligands (compare Figure 4.3b.1,2 to Figure 4.3a.1,2). On the other hand, for the lowest PWR cases, we observed an almost complete depletion of water molecules near $z = 0$. A similar water depletion effect was observed in the hourglass (+) BB simulations in all cases except for the highest PWR. Figure 3b also shows an asymmetric depletion of water molecules for all (+) BB geometries, with lower density values in the upstream BB region ($z > 0$Å). Ligands inside any of the BB geometries did not form any kind of an ordered structure as characterized by the lack of well-defined periodic peaks (Figure 4.3c). In contrast to the trends observed for water molecules, we observed that ligand density tended to decrease as the PWR increased in all BBs (Figure 4.3c).

4.4.2 Radial Distribution Functions

To further characterize the effects of water structuring on the transport of ligands, we calculated the oxygen (O) radial distribution function (RDF) relative to all confined oxygen, and hydrogen (H) atoms (g_{OO}, and g_{OH}, respectively). Due to confinement

effects induced by BB geometry, direct comparison of RDFs between confined and bulk water molecules is not possible without using a geometry correction factor. Therefore, the calculated g_{OO}, and g_{OH} distributions were corrected using a method offered by De Marzio et. al.,[177] Figure 4.4 shows the RDF for oxygen-oxygen pairs confined inside the three BB geometries at varying PWRs in (+) simulations. In addition, we have calculated the time-averaged number of hydrogen bonds (<HB>) formed by the water molecules confined inside the three BB geometries at varying PWRs for the (–/+) simulations (Figure 4.4). <HB> was calculated using the geometric criteria of $r_{OO} < 3.3$Å, and $\phi_{HOO} < 30^o$, where r_{OO} is the distance between the donor and acceptor oxygen atoms, and ϕ_{HOO} is the angle between the intramolecular O-H bond and r_{OO}. All calculated <HB> values were lower than the value of 3.75 reported by Nieto-Draghi for SPC/E bulk water.[178] Furthermore, the maximum value of <HB> for all (–/+) BB was obtained for the highest PWR in the cylindrical (–) BB and was calculated to be 3.52. The cylindrical BB at the highest PWR also showed the smallest ligand effect on <HB> (3.52, vs. 3.45) for the (–/+) cases, respectively.

Table 4.2 lists the heights of the first and second RDF peaks for the three BB geometries at varying PWRs in the (–/+) simulations. The position of the first maxima for g_{OO} was at 2.75Å, slightly shifted to the left from the reported value of 2.78Å.[179] On the other hand, the distances measured for the second maxima and the first minima were within 0.01Å of the values reported in the literature.[179] We have also calculated the height of the peaks in g_{OO}. For all PWRs in (–/+) BBs the first maxima was sharper and larger than the computed bulk value of 3.05 for SPC water molecules.[179] The first maxima for all (+) BBs were characterized by a sharp

decrease in height with a decrease in the PWR. In contrast, the second maxima did not show any large variation in height with respect to the PWRs for the any (–) BB simulations, and the height was within 15% of the previously reported results.[179] The largest variation in height of the second maxima was observed in the barrel (+) simulations as a result of water being depleted from inside the BB due to the transport of ligands.

4.4.3 Thermodynamics and Kinetics of Ligand Transport

To determine the thermodynamic factors governing the transport of ligands along the BBs, we calculated the total ensemble average potential energy ($\langle U \rangle$) inside the building blocks in (–/+) simulations (Figure 4.5). The total average potential energy is comprised of six different pairwise interactions: $\langle U \rangle = \langle U_{WW} + U_{WL} + U_{WC} + U_{LW} + U_{LL} + U_{LC} \rangle$ where the subscripts W, L, and C correspond to water, ligand, and BB wall atoms, respectively. For all (–) simulations $U_{WL} = U_{LW} = U_{LL} = U_{LC} = 0$ since there were no ligands present in these simulations.

Figure 4.5 shows the ensemble average potential energy profile along the radial axis of the -/+ BBs for varying PWRs, where $r = 0$ indicates the centerline of the BB. For all cases, the minimum for ensemble average potential energy ($\langle U \rangle$) was located at the centerline of the BBs. This was the result of water molecules organizing in a way that minimized the number of unfavorable interactions with the BB walls while also maximized the number of HBs. On the other hand, $\langle U \rangle$

approached zero at higher radial distances where the water molecules were in close proximity to the BB walls, and the unfavorable water-wall interactions started to dominate the phenomenon, decreasing the number of possible HBs. Figure 4.5a, also shows that for the (–) barrel and (–) hourglass BBs there is a large jump in U between the lowest and second lowest PWRs. This large decrease in U indicates that number of favorable water-water interactions inside that BB configuration decreased due to smaller number of water molecules found inside the BB, i.e. a result of the low density regions observed in Figure 4.3a. Figure 4.5b also shows that the difference in $\langle \mathrm{U} \rangle$ for the (–/+) cylindrical BBs at the highest PWR is minimal. However, as PWR decreases the difference in $\langle \mathrm{U} \rangle$ between the (–/+) BBs start to become noticeable. This is due to the decrease in favorable water-water interactions and increase in the number of unfavorable water-ligand interactions.

Tables 4.3, 4.4 and 4.5 show the ensemble potential energy difference ($\langle \Delta U \rangle$), excess entropy (S_x), and ensemble average Gibbs free energy difference ($\langle \Delta G \rangle$), where Δ is the difference in the value of the property between the calculated values for the (+) and (–) simulations, i.e., $(\langle \Delta U \rangle) = U_+ - U_-$. Our results show that replacing a water molecule by a ligand, inside the BB, leads to an enthalpic gain for all PWRs in the three tested BB geometries. This was expected as displacing a water molecule from inside the BB to the bulk reservoir frees it from its confined state where the hydrogen bond network is "energetically frustrated."[180] In addition, we found that the level of energetic confinement inside the BB is primarily affected by the PWR, as observed by the decrease of enthalpic gains with respect to an increasing PWR. Next, we calculated the excess entropy of water inside the BBs. This property pro-

vides a quantitative measure of the structural correlations induced by the non-bonded interactions that exist between the water molecules. It also provides a qualitative description of the molecular mobility within the BBs as it has been extensively shown[181] that stronger structural correlations (i.e. higher $-TS_x$ values) lead to slower water dynamics and slower self-diffusion coefficients. The results shown in Tables 4.3, 4.4 and 4.5 (i.e. TS_x vs. D_{W}) show that the structural correlations between water molecules increase as the PWR increases for all BB geometries leading to slower self-diffusion coefficients of water (D_{W}). Furthermore, we found that when comparing $-TS_x$ at all PWR for the three BB geometries the values always followed the pattern barrel > cylinder > hourglass, with the strongest correlations of all cases observed in the barrel BB at the highest PWR yielding a value of $-TS_x = 3.86$ kcal/mol. This is the closest value to the calculated excess entropy of a SPC bulk water molecule, $-TS_{x,\mathrm{calc}} = 4.17$ kcal/mol.[182] Finally, based on the changes in internal energy and excess entropy, we calculated the ensemble average change in Gibbs free energy ($\langle \Delta G \rangle$). As shown in Tables 4.3, 4.4 and 4.5, $\langle \Delta G \rangle$ is negative for most of the tested conditions, except for the three highest PWRs of the barrel BB. In these three specific cases the enthalpic gain of replacing the water molecule inside the BB with a ligand was not high enough to overcome the entropic contributions. Finally, the calculated $\langle \Delta G \rangle$ values at PWRs followed a pattern similar to the one observed for S_x.

In addition to the thermodynamic properties, we also calculated the self-diffusion coefficient of water ($\mathrm{D_W}$), and ligands ($\mathrm{D_L}$) in BBs of different PWRs as shown in Tables 4.3, 4.4 and 4.5. Cylindrical diffusion coefficients (i.e. radial, axial and tangential) are given in Tables 4.9, 4.10 and 4.11. Water diffusion in the

cylindrical BB with a PWR = 0.8 ($D_W = 2.64 \times 10^{-5}$ cm/s) is very close to what is observed in the bulk ($D_W = 2.59 \times 10^{-5}$ cm/s).[174] At equivalent PWR values, Dw always followed the pattern; hourglass > cylinder > barrel. The slowest ligand diffusion was in the cylindrical BB with the highest PWR, reaching a value $D_L = 1.37 \times 10^{-5}$ cm/s. Experimentally measured diffusion coefficient of methane in bulk water is $D_L = 1.82 \times 10^{-5}$ cm/s.[183] In all other BBs an enhancement in D_L – as compared to the bulk – was observed; this is in agreement with the observations made in carbon,[148] and silica[184] nanotubes. The hourglass BB showed a decrease in D_L as the PWR increased, but the D_L range between the lowest and highest PWR was not as large as it was for the cylinder and barrel BBs. The fastest D_L in all BB geometries were measured at the lowest PWRs, with the barrel BB accelerating ligand transport the most, providing an enhancement of over six times as compared to the bulk.

In addition to the thermodynamic properties listed in Tables 4.3, 4.4 and 4.5 we have calculated kinetics of ligand transport along the BBs as shown in Tables 4.6, 4.7 and 4.8. The times listed in Tables 4.6, 4.7 and 4.8 provide information about the behavior of the ligands during transport through the BB. During each transport event, at all PWRs for all BB geometries, ligands may spend some time absorbed to the wall of the BB ($\bar{t}_{ads}$). The fraction of time that ligands spent adsorbed onto the BB wall during each transport event is listed in the third column ($\bar{t}_{frac}$). $\bar{t}_{frac}$ correlates with the path of ligand transport along the BB: A lower $\bar{t}_{frac}$ value implies that the ligand followed a path closer to the centerline of the BB (more data on ligand preferred paths can be found in Section 4.4.4). The last three columns provide information about the

average residence time of ligands depending on their entrance/exit location to the BB. The average residence times are split into three different categories: i) ligands that enter and exit at $z = 10$ (i.e. upstream); ii) ligands that enter and exit at $z = -10$ (i.e. downstream); and iii) ligands that enter at $z = 10$ and exit at $z = -10$ (i.e. full transport along the entire BB). As expected for all three BB geometries the $10 \Rightarrow -10$ case is the longest as the ligand has to traverse the entire BB.

4.4.4 Ligand Transport Pathways

We constructed contour maps of ligand density in order to determine the most preferred paths the ligands followed during transport inside the BBs (Figure 4.6 a-c.1). We defined the condition of high ligand density as the initial concentration of ligands in the upper reservoir ($7.4 \times 10^{-4}\text{Å}^{-3}$) normalized by the bulk density of water ($3.336 \times 10^{-2}\text{Å}^{-3}$), i.e., $\rho_{\text{L}}/\rho_{\text{bulk}} > 0.022$. The ligand contour maps (thick black lines) are overlaid on top of water density heat maps to compare the regions of high water density and high ligand density. In the low PWR cylindrical BB, ligands were primarily transported along the water exclusion zone near the wall. This mode of transport is consistent with the adsorption results presented in Tables 4.3, 4.4 and 4.5. When the PWR of the cylinder BB increased, the preferred paths shifted towards the centerline. Similar results were also observed for the barrel and hourglass BBs.

We also constructed potential energy surfaces (PES) to determine the most energetically favorable paths for the ligands. The PES were constructed using the

weighted histogram analysis method,[185] using the axial and radial directions as the principal analysis axes. The PES are shown as grayscale heat maps in Figure 4.6 a-c.2, where the darker regions correspond to energetically favorable locations inside the BBs. We then constructed hypothetical paths that could be traversed by the ligands from the upper to lower reservoirs. We defined a total of 600 starting points set at $z = 12$Å in the range $0 < r < 6$ at 0.01Å intervals. A similar number of end points was set at $z = -12$Å over the same interval, resulting in 360000 possible start/end point combinations. For all start/end points, we used Dijkstra's algorithm to calculate the path that minimized the sum of the energy gradients in the PES. These paths are shown as the white lines in Figure 4.6 a-c.2 while thicker lines indicate a higher probability for a given path. The results from the energy landscape analysis show that in the BBs at lower PWRs, the ligands preferred to move closer to the walls. This region corresponds to the water exclusion layer (compare location of white lines in Figure 4.6 a-c.2 to water density maps in Figure 4.6 a-c.1). Based solely on an energetic analysis, we observe that the preferred path for the ligands inside the barrel BB at PWR= 0.6 was near the BB wall. This is because the energy gradients are almost non-existent since there is a very low water density inside this BB. Therefore, the ligands are free to move without any hindrance from water molecules.

4.5 Discussion

Broad-substrate specificity enzymes control transport of substrates from the bulk environment into the buried active site, primarily through hydrophobic channels.[19]

There is crystallographic,[13] and computational[119] evidence that these enzyme channels promote formation of low density single-file water networks within. However, the ways in which water networking inside these channels affects ligand transport remains largely undetermined. To tackle this issue, we built a set of coarse-grained building blocks that allowed us to study the behaviors of the confined water molecules to establish their role in ligand transport through hydrophobic channels. The BBs were chosen to represent the most frequently observed features of enzyme channels; cylinders, barrel, and hourglass shapes.[119] The hydrophobic walls of enzyme channels[24] was simulated by varying the non-bonded interaction potential strength of the BBs.

We observed that different BB geometries at equal Lennard-Jones potential well depths, ε_{C}, had different "effective" hydrophobicities with hourglass $<$ cylinder $<$ barrel. This shows showed that the value of ε_{C} cannot be used as the sole parameter to compare hydrophobicity levels of cylindrical vs. non-cylindrical channels. In terms of real BSS enzymes channels, it means that the same type of amino acid can have impose different levels of hydrophobicity levels depending on the wall surface geometry of the surface it is forming. This provides a possible explanation to for the a wide range of hydrophobicity levels that are observed inside BSS enzymes channels[186,24] despite the fact that there are only nine hydrophobic amino acids.

We found that the "effective" hydrophobicity of the BB surface directly affects ligand transport along the channel. For instance, in the barrel BB at the lowest PWR, the formation of a large water exclusion zone adjacent to the wall provides a

region of "low apparent viscosity" where ligands can move faster than in the bulk. Formation of such a large water exclusion zone is due to depletion of HBs induced by surface curvature effects. A similar decrease in HB have been proposed to be the reason for acceleration of water molecules near the surface of cylindrical CNTs.[174] Our results demonstrate that ligands might be moving at different speeds inside BSS enzyme channels depending on their hydrophobic environment. Currently, there is no experimental evidence for this behavior inside BSS enzyme channels, but with recent improvements in the resolution of time-resolved crystallography methods, this hypothesis might possibly be tested in near future.

Ligand diffusion coefficients (D_L) are negatively correlated to residence times $\bar{t}_{res}$. This was observed in the barrel BB because there are single water molecules capping the entrance and exit regions. This means that for a ligand to enter the barrel BB it must displace the capping water molecule, since $\sigma_W \approx r_i$. Displacing these water molecules require that the hydrogen bond network is broken, which is energetically and entropically unfavorable. We believe that in real enzymes with barrel-shaped active sites, these capping water molecules temporarily blocks the ligand from exiting the active site into the transport channel. This gives the ligand enough time to explore different configurations until it finds the proper position that initiates catalytic activity. If the ligand did not have this time to explore the interior of the active site it would be transported back to the bulk before the catalytic reaction starts.

Displacing the capping water molecule in the hourglass bottleneck is more favorable than displacing the capping waters in the barrel BB. This is because the hour-

glass provides a lower "effective" hydrophobicity that allows more water molecules to fit into the interior than the barrel BB. Therefore, waters from either side of the bottleneck can replace the displaced water more readily than in the analogous barrel BB case (note that hourglass r_o = barrel r_i). The implication is that ligands spend less time inside the hourglass BB than in the barrel BB, allowing them to reach the lower reservoir faster. This is counterintuitive as it would be expected that the bottleneck in a real enzyme should block the passage of ligands that could become potentially harmful to the cell if they are reacted upon.[187] Nonetheless, this short residence time effect also makes ligands move back to the upper reservoir fast if they don't overcome the bottleneck barrier. This observation provides a possible explanation as to why BSS enzymes have evolved to have bottleneck-like components as part of their channel geometry. The reduced hydrophobicity and water exclusion zone in the bottleneck decreases the chance of a non-substrate strongly adsorbing onto the channel walls. The fast movement of ligands (both upstream and downstream) prevents the obstruction of what is frequently the only entrance route to the catalytic center.

We propose three possible expansion routes for our building block model. First, altering the PWR of the ligand would affect its transport properties. A lower PWR would displace more water molecules from the BB interior. The increase in water exclusion zone increase D_L. Second, ligands of non-spherical shapes can be considered. For instance, an ellipsoid would be more suitable to study the transport properties of large planar compounds. Alternatively, two bonded spheres to represent a single ligand would allow us to model substrates with asymmetrical hydrophobicity. Third, incorporating hydrophilic effects. The presence of a polar group in BB walls, or on the

ligand, would considerably change the way that surrounding water molecules behave. Raghunathan showed that the flow of water in cylindrical pores is affected more by negative than positive charges.[188] All of these modifications would affect the transport properties of ligands and allow us to expand the scope of our building blocks.

This study led to a better understanding of the dynamics of water and ligands inside hydrophobic biological channels. These observations are relevant phenomena applicable to $\sim$ 3500 different BSS enzymes. The BBs and thermodynamic results presented have been used to develop a non-dimensional model that can predict the transport of ligands through nanochannels found in broad-substrate specificity enzymes. The accuracy of the model, published elsewhere, is 90% and it is up to six orders of magnitude faster than all-atom methods.[131]

4.6 Conclusion

We present a systematic analysis of water networking inside non-cylindrical hydrophobic structures in order to elucidate the role water hydrogen bonding plays on ligand transport. We established that hydrophobic confinement promotes formation of a water exclusion zone adjacent to the walls, and this effect is influenced by the wall geometry such that a concave channel amplifies the hydrophobic exclusion effect as compared to cylindrical or convex geometries of the same non-bonded interaction strength. Within the water exclusion zone, frictional resistance is reduced, significantly accelerating the hydrophobic ligand transport across. Accelerated hydrophobic

ligand transport through hydrophobic geometries is also explained from a thermodynamic perspective of disrupting water hydrogen bonding network. At one extreme, we found that the entropic contributions to the structural order of water inside the geometry might preclude the ligand from entering, as this would cause a disruption on the fragile hydrogen bond network of the confined and energetically "frustrated" water molecules. On the other extreme, if the enthalpic contributions are very high, the water molecules are displaced to the bulk as the small hydrophobic ligand adsorbs onto the hydrophobic wall, slowing down diffusion. The results presented here, in particular for the highest hydrophobicity cases, are consistent with the observations made in crystallographic,[13] and computational[119] studies. The BB geometries and the thermodynamic results presented, have successfully been used elsewhere to develop a model that can predict the transport of ligands through nanochannels found in broad-substrate specificity enzymes.[131]

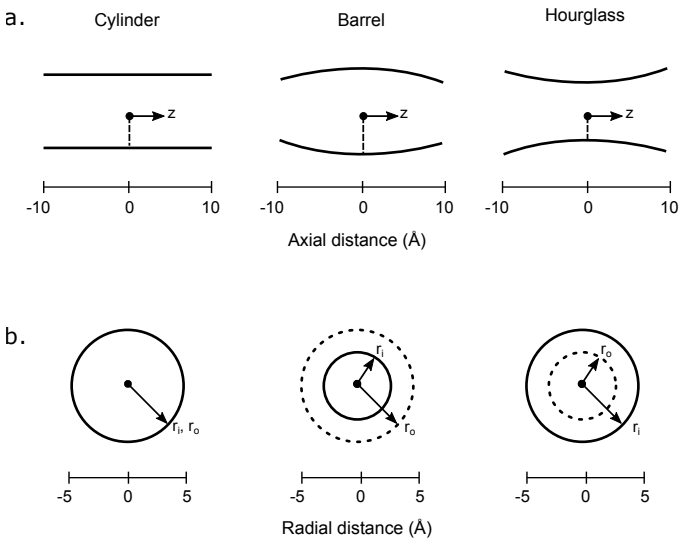

Figure 4.1: Coordinate projection of the system showing the three type of building blocks used in this study. **a.** Projection along the axial coordinate (z-axis), and **b.** Projection along the radial coordinate showing the cylinder, barrel and hourglass geometries. The dashed lines correspond to the radius (r_i) of the channel at z = 0 and the solid lines mark the radius (r_o) at z = $\pm$10Å.

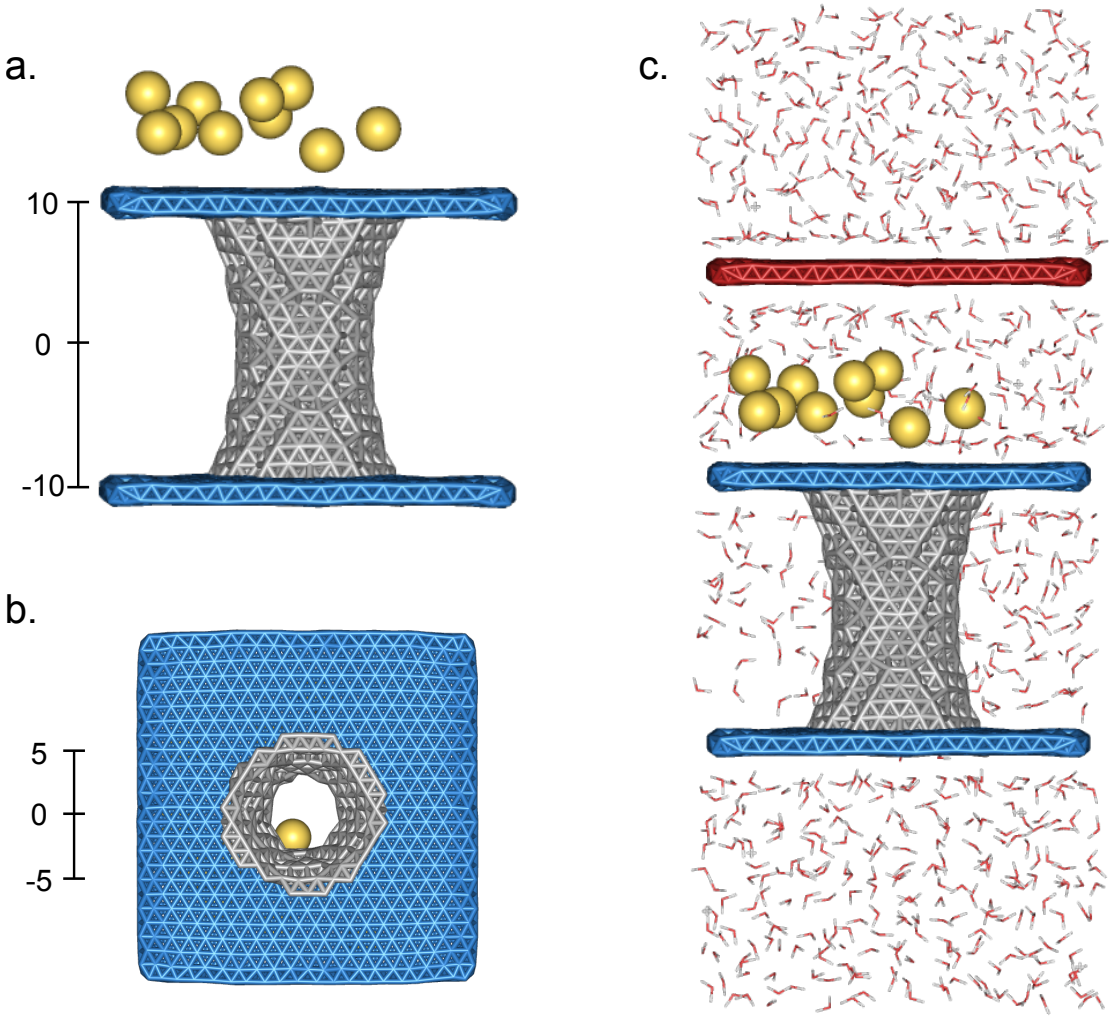

Figure 4.2: Simulation system. (a) Side view showing the hourglass building block in grey, the ligands (shown as yellow spheres) are shown at their initial position in the upper bulk reservoir. (b) Bottom view of the hourglass building block. (c) The entire system showing the top and bottom water reservoirs. Note that the water molecules above the red plate are still considered to be in the lower reservoir due to the periodic boundary condition applied.

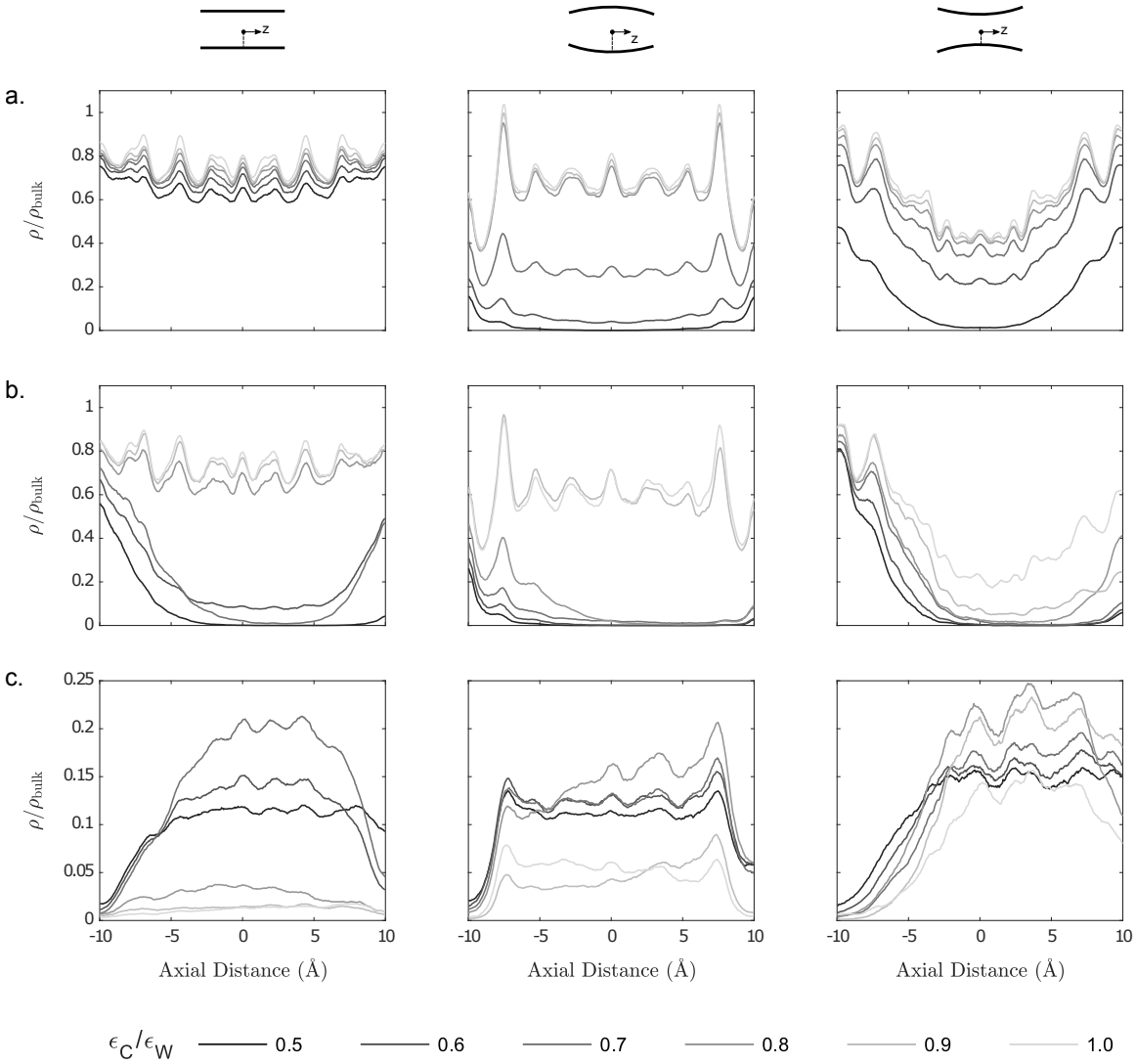

Figure 4.3: Time averaged density profiles along the axial direction of the BBs. Density profile of water inside the BB for simulations (a) without ligands, and (b) in the presence of hydrophobic ligands. (c) Density profile of ligands inside channels. Note that ligand diffusion is from right to left. The left column is for the cylindrical BB, center column is for the barrel-shaped BB, and the right column is for the hourglass-shaped BB. Darker lines correspond to profiles obtained with increasingly more hydrophobic walls.

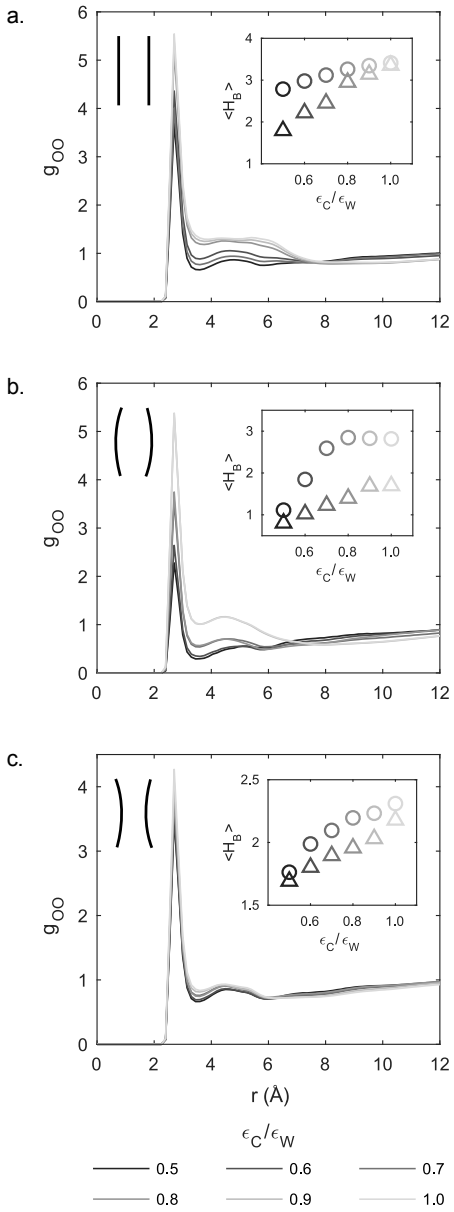

Figure 4.4: Radial distribution function (RDF) for water O-O interactions. RDF for the (a) cylindrical, (b) barrel, and (c) hourglass BBs. All RDF plots correspond to the (+) simulations, i.e. containing water and ligands. Darker lines correspond to profiles obtained within increasingly more hydrophobic walls. The insets show the average number of hydrogen bonds formed per water molecule inside the BB. The circles correspond to (–) simulations, and the triangles correspond to (+) simulations containing water and ligands.

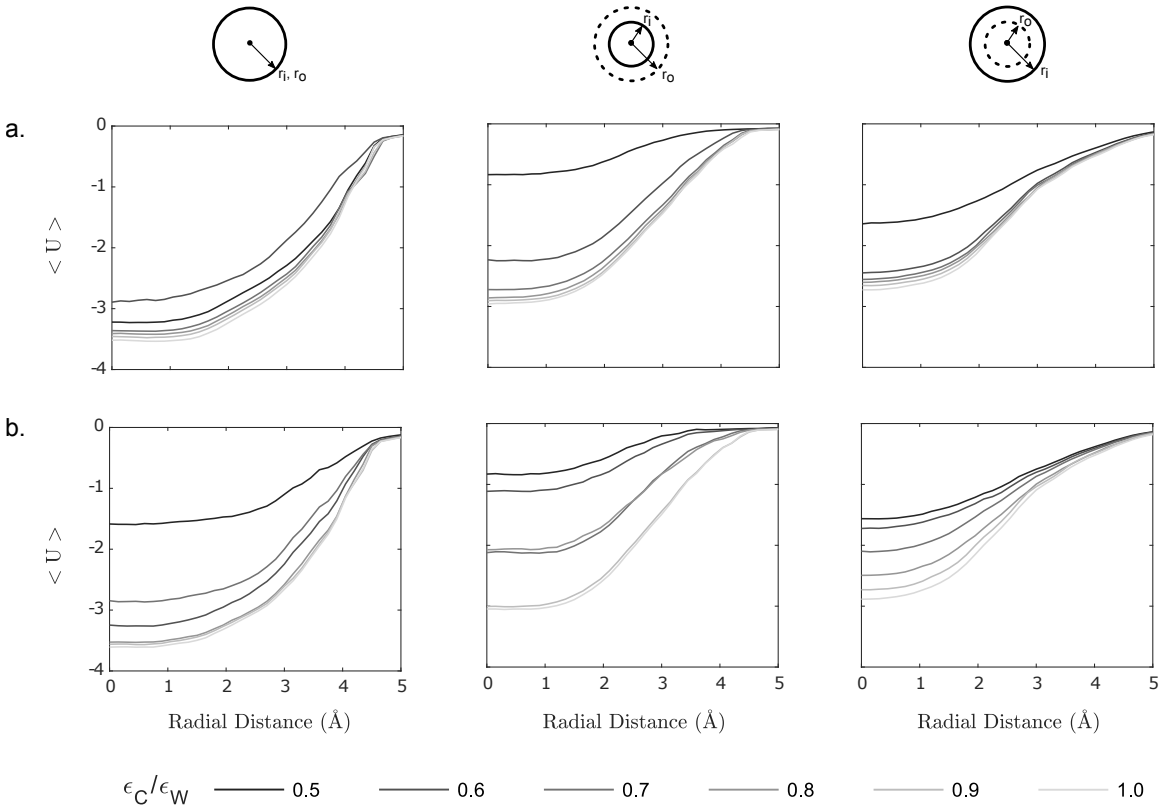

Figure 4.5: Ensemble average energy profile along the radial axis for (–/+) simulations. The total internal energy includes contributions from all atom pair interactions, i.e. water, ligands (if present in simulation), and the BB walls. (a) Internal energy profiles inside BBs flooded with water; (–) simulations. (b) Internal energy profiles inside BBs for simulations containing water and ligands; (+) simulations. The left column is for the cylindrical BB, center column is for the barrel-shaped BB, and the right column is for the hourglass-shaped BB. Darker lines correspond to profiles obtained with increasingly more hydrophobic walls.

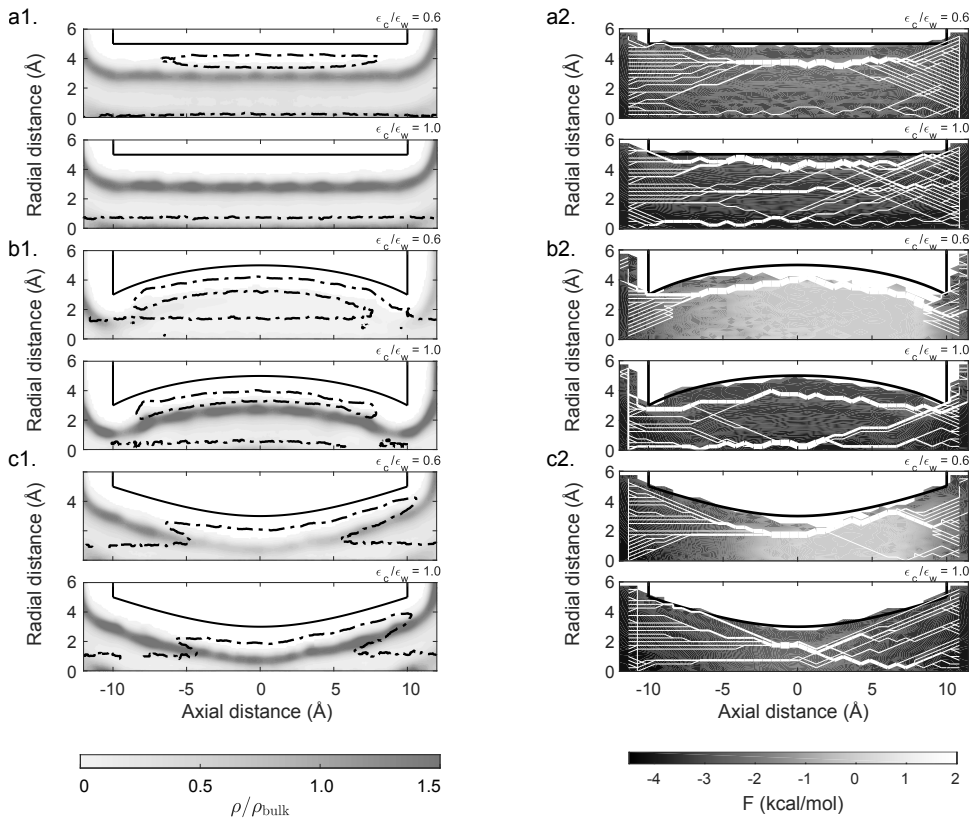

Figure 4.6: (a-c.1) Maps showing water density inside the three BB geometries at low and high PWR ($\varepsilon_C/\varepsilon_W$ = 0.6 and 1.0). The darker lighter regions of the heat map indicate low water density (ρ_W/ρ_{bulk}), and lighter darker regions indicate ρ_W close to bulk values. White regions indicate the lack of water molecules at that location. The area bound by the thick blackdot-dash contour lines shows regions of high ligand density regions. We defined the condition of high ligand density as the initial concentration of ligands in the upper reservoir (7.4×10^{-4}Å^{-3}) normalized by the bulk density of water (3.336×10^{-2}Å^{-3}), i.e., $\rho_L/\rho_{bulk} > 0.022$. The thick black lines show the location of each BB wall. Transport of ligands occurs from right to left. (a-c.2) Potential Energy Surface (PES) maps showing the energetic barrier needed to transport a ligand along the axial and radial directions at low and high PWR, ($\varepsilon_C/\varepsilon_W$ = 0.6 and 1.0). The darker regions of the heat map indicate a more favorable free energy (F), whereas the lighter regions indicate a less favorable free energy. The white lines show hypothetical paths that a ligand can take while transported downstream, from $z = 12$Å to $z = -12$Å. Thicker lines indicate a higher probability that a ligand would follow that path. The thick black lines show the location of each BB wall .

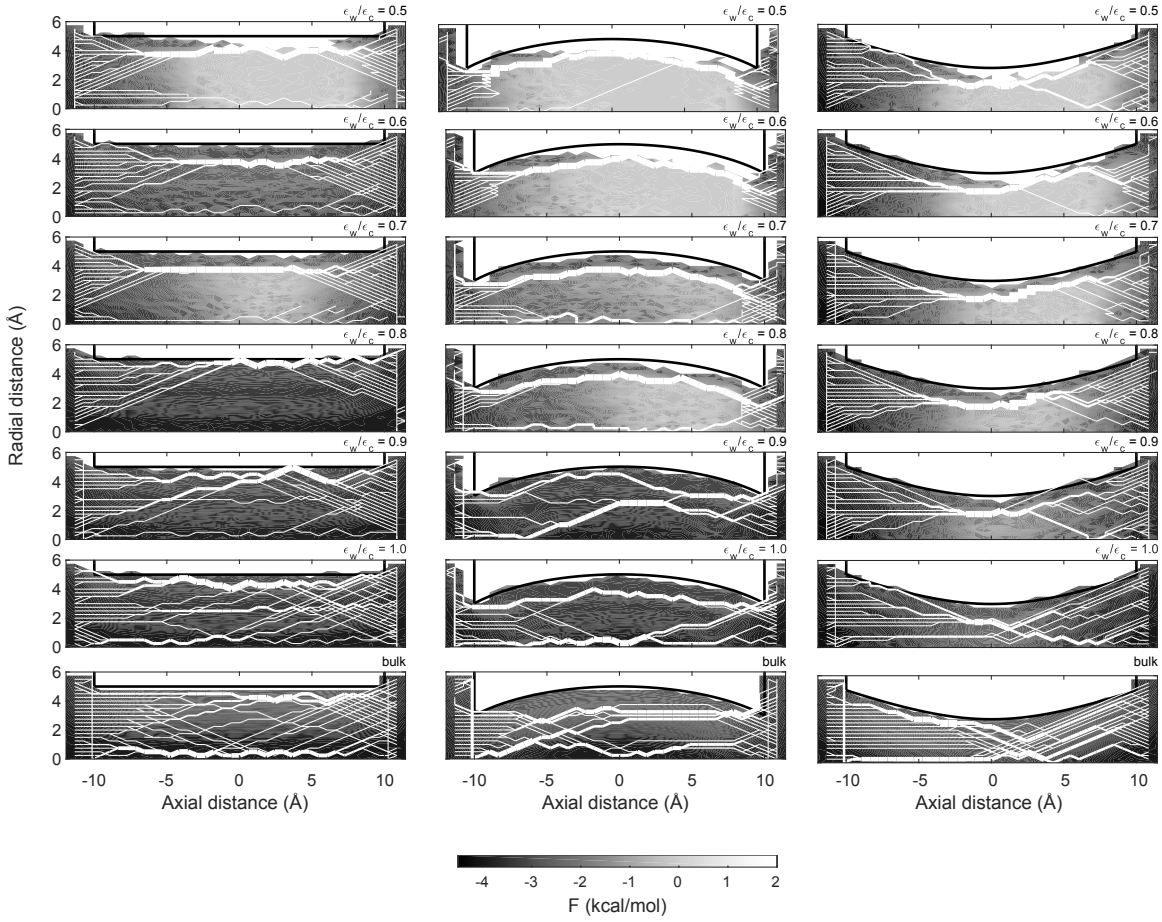

Figure 4.7: Potential Energy Surface (PES) maps showing the energetic barrier needed to transport a ligand along the axial and radial directions in the BB at all PWR, $\varepsilon_{\mathrm{W}}/\varepsilon_{\mathrm{C}} = 0.5 - 1.0$, respectively. (a) cylinder BB, (b) barrel BB and, (c) hourglass BB. The white lines show the different paths taken by the ligands while transported downstream, from $z = 12$Å to $z = -12$Å. Thicker lines indicate a higher probability that a ligand would follow that path.

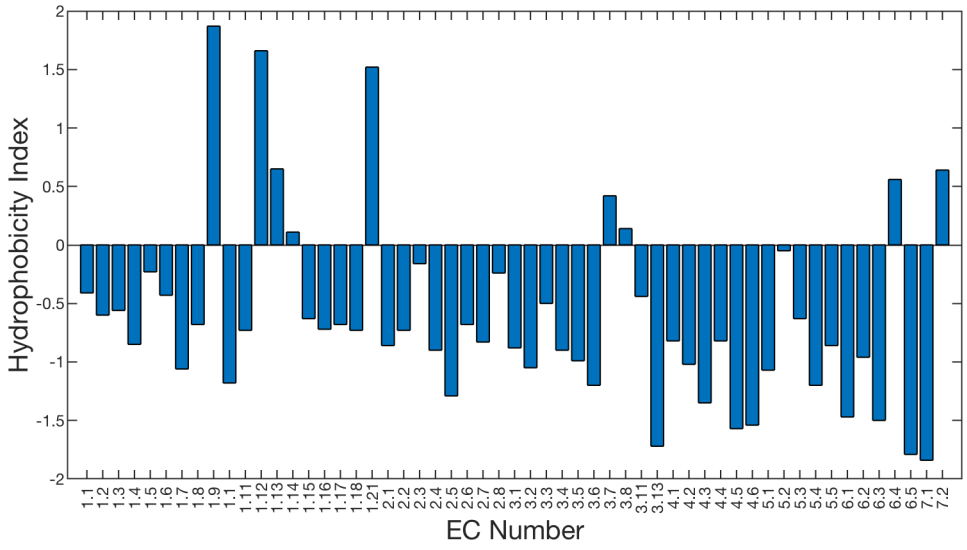

Figure 4.8: Hydrophobicity index of channels found in enzymes belonging to each enzyme commission (EC) subclass. A positive value indicates the presence of primarily hydrophobic residues inside the channel, whereas a negative value indicates the presence of primarily hydrophilic residues inside the channel.

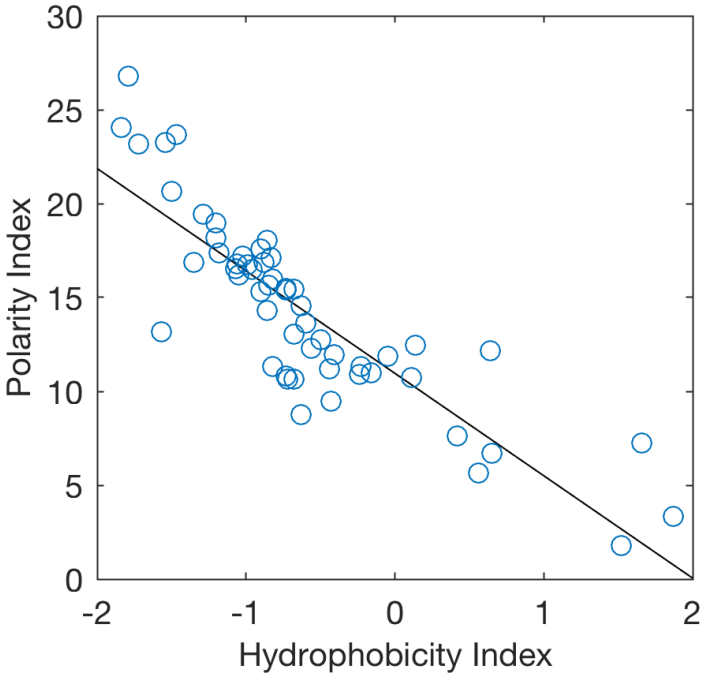

Figure 4.9: Relationship between polarity index and hydrophobicity index for residues found inside channels for enzymes belonging to all EC subclasses. A hydrophobic index positive value indicates the presence of primarily hydrophobic residues inside the channel, whereas a negative value indicates the presence of primarily hydrophilic residues inside the channel. Lower polarity index values indicate the absence of polar residues, i.e. less chances of forming hydrogen bonds with water or ligands traversing the channel.

Table 4.1: Geometric parameters used for the three building blocks (Figure 4.1), along with the force field parameters used to describe the SPC water molecules based on the OPLS force field.[43]

	Geometric Parameters		
	Cylinder	**Barrel**	**Hourglass**
r_i (Å)	5	3	5
r_o (Å)	5	5	3

	Force Field Parameters		
Atom	***q* (a.u)**	**σ(Å)**	**ε (kcal/mol)**
O	-0.82	3.166	0.1554
H	0.41	0	0

Table 4.2: Oxygen-Oxygen pair distribution functions for all BBs and PWRs. All values correspond to the height of the g_{OO} peak. (–) columns correspond to simulations in BBs flooded with water, and (+) columns correspond to the effects induced by the presence and transport of ligands along the BB. [a] Calculated values of g_{OO} for the bulk water using the SPC model based on MD simulations by Mark et. al.[179]

	Cylinder				**Barrel**				**Hourglass**			
$\frac{\varepsilon_C}{\varepsilon_W}$	**1st Max**		**2nd Max**		**1st Max**		**2nd Max**		**1st Max**		**2nd Max**	
	(−)	(+)	(−)	(+)	(−)	(+)	(−)	(+)	(−)	(+)	(−)	(+)
0.5	4.81	3.71	1.21	0.87	3.14	2.29	1.01	0.56	4.08	3.50	0.97	0.85
0.6	4.87	4.03	1.22	0.94	4.05	2.65	1.01	0.60	4.15	3.57	0.92	0.86
0.7	4.93	4.39	1.22	1.23	4.56	3.76	1.05	0.70	4.22	3.85	0.92	0.91
0.8	4.95	5.22	1.22	1.23	4.56	3.76	1.05	0.70	4.32	3.86	0.94	0.91
0.9	5.05	5.42	1.21	1.29	4.66	5.39	1.07	1.16	4.36	4.08	0.94	0.92
1.0	5.12	5.58	1.22	1.32	4.70	5.39	1.08	1.16	4.48	4.28	0.96	0.93
Ref[a]	3.05	-	1.04	-	3.05	-	1.04	-	3.05	-	1.04	-

Table 4.3: Thermodynamic properties, and transport parameters for the cylinder BB. Thermodynamic properties ($\langle \Delta U \rangle$, TS_x and $\langle \Delta G \rangle$) are listed in kcal/mol, and are calculated as the difference between the (+) and (–) simulations. The diffusion coefficients for ligands (D_L) and water (D_W) are in cm^2/s. The volume fraction (V^*) is the volume occupied by the water molecules inside the BB for the (–) simulations normalized to the total volume of the BB.

Cylinder						
$\varepsilon_C/\varepsilon_W$	$\langle \Delta U \rangle$	$-TS_x$	$\langle \Delta G \rangle$	D_L ($\times 10^{-5}$)	D_W ($\times 10^{-5}$)	V^*
0.5	-3.52	1.91	-1.25	7.82	2.86	0.544
0.6	-3.23	2.16	-1.08	7.29	2.76	0.541
0.7	-3.26	2.37	-0.90	6.44	2.68	0.543
0.8	-3.21	2.70	-0.52	1.99	2.64	0.541
0.9	-3.15	2.76	-0.38	1.44	2.62	0.534
1.0	-3.09	2.87	-0.22	1.37	2.46	0.538

Table 4.4: Thermodynamic properties, and transport parameters for the barrel BB. Thermodynamic properties ($\langle \Delta U \rangle$, TS_x and $\langle \Delta G \rangle$) are listed in kcal/mol, and are calculated as the difference between the (+) and (–) simulations. The diffusion coefficients for ligands (D_L) and water (D_W) are in cm^2/s. The volume fraction (V^*) is the volume occupied by the water molecules inside the BB for the (–) simulations normalized to the total volume of the BB.

Barrel						
$\varepsilon_C/\varepsilon_W$	$\langle \Delta U \rangle$	$-TS_x$	$\langle \Delta G \rangle$	D_L ($\times 10^{-5}$)	D_W ($\times 10^{-5}$)	V^*
0.5	-3.49	2.46	-1.03	8.59	2.37	0.359
0.6	-3.28	2.66	-0.72	8.00	1.92	0.409
0.7	-3.13	2.86	-0.27	8.12	1.59	0.425
0.8	-2.97	3.25	0.07	7.62	1.37	0.431
0.9	-2.98	3.56	0.27	3.19	1.39	0.430
1.0	-2.90	3.86	0.54	2.96	1.34	0.428

Table 4.5: Thermodynamic properties, and transport parameters for the hourglass BB. Thermodynamic properties ($\langle \Delta U \rangle$, TS_x and $\langle \Delta G \rangle$) are listed in kcal/mol, and are calculated as the difference between the (+) and (–) simulations. The diffusion coefficients for ligands (D_L) and water (D_W) are in cm^2/s. The volume fraction (V^*) is the volume occupied by the water molecules inside the BB for the (–) simulations normalized to the total volume of the BB.

Hourglass						
$\varepsilon_C/\varepsilon_W$	$\langle \Delta U \rangle$	$-TS_x$	$\langle \Delta G \rangle$	D_L ($\times 10^{-5}$)	D_W ($\times 10^{-5}$)	V^*
0.5	-3.21	1.42	-1.76	4.91	4.45	0.509
0.6	-2.99	1.51	-1.47	4.68	3.35	0.522
0.7	-2.92	1.61	-1.31	4.57	3.15	0.524
0.8	-2.71	1.72	-0.98	4.06	3.04	0.521
0.9	-2.70	1.82	-0.85	4.06	2.89	0.520
1.0	-2.68	2.09	-0.59	4.01	2.77	0.516

Table 4.6: Kinetics of ligand transport along the cylinder BBs. Average ligand residence times ($\bar{t}_{\text{res}}$) and average ligand adsorption times ($\bar{t}_{\text{ads}}$) are in picoseconds, $\bar{t}_{\text{frac}}$ is the fraction of time the ligand is adsorbed to the wall ($\bar{t}_{\text{res}}/\bar{t}_{\text{ads}}$). The last three columns show the average residence time of the ligands inside the BB depending on the ligand entrance/exit position; these times are listed in picoseconds. The subscripts indicate the entrance and exit locations of the ligand i.e., $10 \Rightarrow -10$ means that the ligand entered the BB at $z = 10$ and exited at $z = -10$. All times are calculated based on the (+) simulations.

Cylinder						
$\varepsilon_{\text{C}}/\varepsilon_{\text{W}}$	$\bar{t}_{\text{res}}$	$\bar{t}_{\text{ads}}$	$\bar{t}_{\text{frac}}$	$\bar{t}_{10\Rightarrow 10}$	$\bar{t}_{-10\Rightarrow -10}$	$\bar{t}_{10\Rightarrow -10}$
0.5	21	18	0.86	16	14	26
0.6	52	40	0.77	36	32	63
0.7	67	35	0.52	39	42	74
0.8	78	25	0.32	49	58	100
0.9	90	15	0.17	93	97	107
1.0	94	10	0.11	94	98	118

Table 4.7: Kinetics of ligand transport along the barrel BBs. Average ligand residence times ($\bar{t}_{\text{res}}$) and average ligand adsorption times ($\bar{t}_{\text{ads}}$) are in picoseconds, $\bar{t}_{\text{frac}}$ is the fraction of time the ligand is adsorbed to the wall ($\bar{t}_{\text{res}}/\bar{t}_{\text{ads}}$). The last three columns show the average residence time of the ligands inside the BB depending on the ligand entrance/exit position; these times are listed in picoseconds. The subscripts indicate the entrance and exit locations of the ligand i.e., $10 \Rightarrow -10$ means that the ligand entered the BB at $z = 10$ and exited at $z = -10$. All times are calculated based on the (+) simulations.

Barrel						
$\varepsilon_{\text{C}}/\varepsilon_{\text{W}}$	$\bar{t}_{\text{res}}$	$\bar{t}_{\text{ads}}$	$\bar{t}_{\text{frac}}$	$\bar{t}_{10\Rightarrow 10}$	$\bar{t}_{-10\Rightarrow -10}$	$\bar{t}_{10\Rightarrow -10}$
0.5	57	49	0.86	55	60	64
0.6	65	55	0.85	63	55	70
0.7	100	60	0.60	97	92	110
0.8	120	65	0.54	94	104	132
0.9	260	150	0.58	168	161	367
1.0	485	190	0.39	380	400	560

Table 4.8: Kinetics of ligand transport along the hourglass BBs. Average ligand residence times ($\bar{t}_{\text{res}}$) and average ligand adsorption times ($\bar{t}_{\text{ads}}$) are in picoseconds, $\bar{t}_{\text{frac}}$ is the fraction of time the ligand is adsorbed to the wall ($\bar{t}_{\text{res}}/\bar{t}_{\text{ads}}$). The last three columns show the average residence time of the ligands inside the BB depending on the ligand entrance/exit position; these times are listed in picoseconds. The subscripts indicate the entrance and exit locations of the ligand i.e., $10 \Rightarrow -10$ means that the ligand entered the BB at $z = 10$ and exited at $z = -10$. All times are calculated based on the (+) simulations.

Hourglass						
$\varepsilon_{\text{C}}/\varepsilon_{\text{W}}$	$\bar{t}_{\text{res}}$	$\bar{t}_{\text{ads}}$	$\bar{t}_{\text{frac}}$	$\bar{t}_{10\Rightarrow 10}$	$\bar{t}_{-10\Rightarrow -10}$	$\bar{t}_{10\Rightarrow -10}$
0.5	9	4	0.44	7	15	20
0.6	10	5	0.50	9	18	27
0.7	11	5	0.45	10	20	30
0.8	13	6	0.46	11	25	49
0.9	15	8	0.53	12	32	50
1.0	18	9	0.50	16	43	60

Table 4.9: Diffusion coefficients for ligands inside the cylinder BB geometry. The radial (D_R) and axial (D_z) diffusion coefficients are in cm^2/s, and the tangential D_θ diffusion coefficients are in deg^2/s.

	Cylinder		
$\varepsilon_C/\varepsilon_W$	D_R ($\times 10^{-5}$)	D_z ($\times 10^{-5}$)	D_θ
0.5	6.29	13.86	31
0.6	5.83	13.51	29
0.7	5.17	11.62	28
0.8	1.56	2.90	18
0.9	1.29	2.41	17
1.0	1.25	2.31	16

Table 4.10: Diffusion coefficients for ligands inside the barrel BB geometry. The radial (D_R) and axial (D_z) diffusion coefficients are in cm^2/s, and the tangential D_θ diffusion coefficients are in deg^2/s.

	Cylinder		
$\varepsilon_C/\varepsilon_W$	D_R ($\times 10^{-5}$)	D_z ($\times 10^{-5}$)	D_θ
0.5	7.08	18.12	30
0.6	6.75	18.08	29
0.7	6.44	16.44	28
0.8	6.09	16.01	26
0.9	3.36	8.31	24
1.0	2.89	7.15	20

Table 4.11: Diffusion coefficients for ligands inside the hourglass BB geometry. The radial (D_R) and axial (D_z) diffusion coefficients are in cm^2/s, and the tangential D_θ diffusion coefficients are in deg^2/s.

	Cylinder		
$\varepsilon_C/\varepsilon_W$	D_R ($\times 10^{-5}$)	D_z ($\times 10^{-5}$)	D_θ
0.5	3.94	8.80	30
0.6	3.84	8.60	24
0.7	3.61	8.01	15
0.8	3.49	7.92	14
0.9	3.46	7.86	13
1.0	3.49	7.81	12

Table 4.12: Subclasses of enzymes containing hydrophobic channels leading into the active site.

E.C Subclass	Description	PDB Entries	Number of Organisms
1.9	Acting on a heme group of donors	557	340
1.12	Acting on hydrogen as donor	268	259
1.13	Acting on single donors with incorporation of molecular oxygen (oxygenases)	640	850
1.14	Acting on paired donors with incorporation or reduction of molecular oxygen	3200	2346
1.21	Acting on X-H and Y-H to form an X-Y bond	65	146
3.7	Acting on carbon-carbon bonds	71	140
3.8	Acting on halide bonds	166	157
6.4	Forming carbon-carbon bonds	122	202

Chapter 5

Prediction of Ligand Transport Along Hydrophobic Enzyme Nanochannels

5.1 Chapter Summary

Buried active sites of enzymes are connected to the bulk solvent through a network of hydrophobic channels. We developed a discretized model that can accurately predict ligand transport along hydrophobic channels up to six orders of magnitude faster than any other existing method. The non-dimensional nature of the model makes it applicable to any hydrophobic channel/ligand combination.

5.2 Introduction

Nanochannels inside enzymes are responsible for precisely controlling the bidirectional transport of small molecules (ligands) between a buried active site and the cytoplasm.[49,24] These nanochannels may also serve more sophisticated functions such as preventing competing side reactions,[24] protecting the enzyme from toxic or unstable intermediates,[187] and selecting for substrates.[24] During the last decade, several software tools were developed for the identification and characterization of nanochannels.[189] However, none of these software were designed to model the transport of ligands through nanochannels and enable rapid determination of whether a ligand is capable of reaching the active site. The lack of such a modeling tool necessitates screening and identification of novel substrates using experimental[190] and computational[54,18,119] approaches that are expensive and time-consuming.

In this communication, we describe a coarse-grained model for prediction of ligand transport inside hydrophobic enzyme nanochannels that is faster than the all-atom[119] and steered molecular dynamics[18] alternatives. To reduce the excessive computational requirement for calculating all pairwise interaction potentials, we perform a simple discretization (slicing) procedure with which a hydrophobic channel inside an enzyme is represented as a sequence of building blocks as shown in Figure 5.1a. Each building block is defined by three parameters (Figure 5.2) to describe its geometry and physicochemical characteristics: i) entrance radius: r_i; ii) midpoint radius: r_o; and iii) the intermolecular nonbonded interaction strength (ε). The nonbonded interaction strength of the building block, ε_C, is defined in terms of the Lennard-Jones potential. Similarly, the ligand is modeled as a sphere of uniform hydrophobicity represented by the nonbonded interaction strength, ε_L. We nondimensionalized the building block geometric parameters (e.g. r_o/r_i); and the nonbonded strengths of the building block, and the ligand with respect to the potential well of a SPC water molecule ($\varepsilon_C/\varepsilon_W$, and $\varepsilon_L/\varepsilon_W$, respectively). In addition, the volume fraction of the building blocks inaccessible to water molecules (i.e. the excluded volume, V_O/V_T) was found to be a critical parameter in modeling the transport of ligands. The dimensional analysis allowed the generation of a unified set of topologies that can describe any given hydrophobic channel section/ligand combination. A more detailed explanation of the building block parameters and the dimensional analysis can be found in the methods section. Excluded volume values for each building block is provided in Table 5.1.

5.3 Computational Methods

5.3.1 Molecular dynamics and Tunnel Identification

We ran two 40 ns molecular dynamic simulation for the unbound structure of naphthalene 1,2-dioxygenase (PDB: 1O7G); details of the setup can be found elsewhere.[119] A total of one hundred snapshots were chosen at random from the last 20 ns of the production simulation, these snapshots represent the collective motion of the enzyme domains as well as the breathing motion of the channel leading into the active site. The channels connecting the solvent exposed area and the mononuclear iron were identified using the program MOLE 2.0.[23] Each of the channels identified by MOLE 2.0 can be described by a set of coordinates detailing the centerline of the path connecting the iron (starting point, $d = 0$) and the bulk solvent. Hereafter, we call each individual point along the path of the channel a *step* and identify it by the subscript k. Similarly, at each step k the program returns the hydrophobicity index (H_I) describing the physicochemical environment of the channel. A similar simulation was performed for the unbound structure of NDO but having placed water molecules inside the channel using the "Solvate Pocket" feature of Desmond.[102]

5.3.2 Conversion of hydrophobicity index

MOLE 2.0 calculates the bulk hydrophobicity character of an enzyme channel using the scale proposed by Cid et al.[191] The value of H_I ranges between -1.2 and 1.85, positive values indicate non-polar hydrophobic regions, and negative values indicate polar and hydrophilic regions. In the development of our building block model we used the Lennard-Jones potential well ratio, normalized with respect to that of a SPC/E water molecule, to numerically describe the hydrophobicity of the building block wall surface. To determine a conversion factor between H_I and the potential well ratio used in our building block model, we correlated the hydrophobicity index with a mass-averaged LJPW for each amino acid. This mass-averaged LJPW value for each type of amino acid was calculated using Equation 5.1:

$$\varepsilon_{\mathrm{AA}} = \frac{\sum_i m_i \varepsilon_i}{\sum_i m_i} \tag{5.1}$$

where i refers to every atom in the amino acid and, m is the mass of the atom, and ε is the LJPW for that atom as parameterized by the OPLS force field.

5.3.3 Discretization of the enzyme channel

To construct our building block model we partitioned the enzyme channels into a set of n characteristic geometries each with a specific hydrophobicity (Figure 5.1a)). We first fitted a piecewise polynomial spline on the radius output from MOLE 2.0

for all points k along the centerline of the channel. The spline was used to calculate the radius – r_o – of the channel at a distance d from the starting point as well as the radius – r_i – at a distance $d \pm 0.2$Å from the starting point. The ratio r_i/r_o determines the type of geometry (i.e. barrel, cylinder or hourglass BB) and follows the same nomenclature used in the development of our building block model (Figure 5.1). Similarly, we fitted a cubic spline on the H_I output from MOLE 2.0 for all points k along the centerline of the channel. The calculated H_I spline was converted to the potential well ratio scale. The potential well ratio of the building block was calculated as the geometric average of the potential well ratio between points r_o and r_i. This procedure was repeated for every snapshot, resulting a total of 100 building block models.

5.3.4 Determination of active site entry

Each ligand (small organic molecule) was modeled as a sphere of uniform hydrophobicity. The radius of the ligand sphere – r_L – was calculated from the surface accessible surface area (A_S) for each molecule, i.e $r_L = \sqrt{A_S/4\pi}$. And the potential well ratio for the ligand sphere was calculated by normalizing the mass averaged Lennard-Jones potential well of the molecule (Equation 5.1) to the value of a SPC water molecule, i.e. $\varepsilon_L/\varepsilon_W$. The calculated characteristic parameters were used to determine ΔG^* for each BB/ligand combination, as described in Section 5.3.5. These values were used to construct a ligand trajectory plot along each of the 100 enzyme channels. If at any point along the ligand trajectory plot $\Delta G^* > 0$, it was determined that the ligand did

not enter the active site. Hence the condition of ligand entry was defined as $\Delta G^* < 0$ for *all* building blocks.

5.3.5 Nondimensinalization of building block model

We constructed a *master plot* that can determine if a ligand is likely to successfully enter, and move along, and finally exit each n building block, based on previous thermodynamic and kinetic results. We used four characteristic nondimensional parameters to describe any building block/ligand combination: i) $\varepsilon_{\mathrm{C}}/\varepsilon_{\mathrm{W}}$; ii) $\varepsilon_{\mathrm{C}}/\varepsilon_{\mathrm{L}}$; iii) $\mathrm{r_o}/\mathrm{r_i}$; and iv) $\mathrm{V_o}/\mathrm{V_T}$. The last term is the excluded volume inside each building block at a given potential well ratio, this value can be directly obtained from Table 5.1. These four parameters were non-linearly fitted against the nondimensionalized Gibbs' free energy (ΔG^*), where $\Delta G^* = \Delta G/\mathrm{k_B T}$.

5.3.6 Nondimensional Gibb's Free Energy

The relationship bteween Gibb's free energy (ΔG) and the equilibrium constant (K_{eq}),

$$\Delta G = -k_B \mathrm{T} \ln(K_{eq}) \tag{5.2}$$

was nondimensionalized resulting in:

$$\Delta G/k_B\mathrm{T} = \Delta G^* = \ln(1/K_{eq}) \tag{5.3}$$

where k_B is the Boltzmann constant, and T is the temperature of the system.

The transport of ligands along the building blocks depend on the geometric and physicochemical parameters, as well as the excluded volume fraction. Therefore, the equilibrium constant of our transport process must also be a function of on these parameters. We expressed the equilibrium constant, K_{eq}, as a function of 25 nondimensional geometric and physicochemical parameters in our system in the following form:

$$\frac{1}{K_{eq}} = f\Big(\alpha^{\mathrm{a}}\beta^{\mathrm{b}}\gamma^{\mathrm{c}}...\omega^{\mathrm{z}}\Big) \tag{5.4}$$

inserting Equation 5.4 into Equation 5.3 results in:

$$\Delta G^* = \ln\big(\alpha^{\mathrm{a}}\beta^{\mathrm{b}}\gamma^{\mathrm{c}}...\omega^{\mathrm{z}}\big) \tag{5.5}$$

We then estimated the value of the exponents a, b, c...z using Matlab's iterative least-squared non-linear regression solver (*nlinfit* function). Majority of the exponents were zero, resulting in the equation:

$$\Delta G^* = 0.55\ln\Big[\Big(\frac{\varepsilon_{\mathrm{C}}}{\varepsilon_{\mathrm{W}}}\Big)^4\Big(\frac{\varepsilon_{\mathrm{C}}}{\varepsilon_{\mathrm{L}}}\Big)^{1.5}\Big(\frac{\mathrm{r_o}}{\mathrm{r_i}}\Big)^3\Big(\frac{\mathrm{V_o}}{\mathrm{V_T}}\Big)\Big] \tag{5.6}$$

or,

$$\Delta G^* = 0.55 \log_{10} \left[\left(\frac{\varepsilon_{\mathrm{C}}}{\varepsilon_{\mathrm{W}}} \right)^4 \left(\frac{\varepsilon_{\mathrm{C}}}{\varepsilon_{\mathrm{L}}} \right)^{1.5} \left(\frac{\mathrm{r_o}}{\mathrm{r_i}} \right)^3 \left(\frac{\mathrm{V_o}}{\mathrm{V_T}} \right) \right] \tag{5.7}$$

The first two terms inside the logarithm represent the energetic contributions to the equilibrium constant, whereas the last two terms represent the entropic contributions due to the transport of the ligand through the building blocks. In a favorable process where a ligand gets transported through the building block $\Delta G^* < 0$. Therefore, $\Delta G^* = 0$ is defined as the cutoff between favorable and unfavorable transport process.

5.4 Results and Discussion

The non-linear regression in Figure 5.1b shows the correlation between the normalized Gibb's free energy of transport ($\Delta G^* = \Delta G / k_B \mathrm{T}$) and the dimensionless parameter that characterizes the contributions of geometry and hydrophobicity of the system, as well as exclusion volume effects inside the building blocks.

$$\Delta G^* = 0.55 \log_{10} \left[\left(\frac{\varepsilon_{\mathrm{C}}}{\varepsilon_{\mathrm{W}}} \right)^4 \left(\frac{\varepsilon_{\mathrm{C}}}{\varepsilon_{\mathrm{L}}} \right)^{1.5} \left(\frac{\mathrm{r_o}}{\mathrm{r_i}} \right)^3 \left(\frac{\mathrm{V_o}}{\mathrm{V_T}} \right) \right] = 0 \tag{5.8}$$

Equation 5.8 defines the cutoff between unsuccessful, and successful transport across the building blocks (Shown in Figure 5.1b as the shaded, and unshaded regions, respectively). The cutoff was based on observations made during our building block

model development, where ligands did not traverse the full length of the building block at certain geometry/hydrophobicity conditions.

To test the applicability of our ligand transport model, we used the enzyme naphthalene 1,2-dioxygenase (NDO) natively expressed in *Pseudomonas putida* NCIB 9816-4. It has been shown theoretically,[42] and experimentally[57] that substrate binding to the buried active site of NDO is necessary for catalysis. Since ligands *must* overcome the geometric and/or energetic barriers imposed by the $\sim$ 17Å long channel to reach the active site,[119] any positive catalytic activity can be used as a proxy for successful ligand transport through the channel. We performed two 40ns MD simulations for the unbound structure of NDO to study the effect of water on the geometry and hydrophobicity of the channel, i.e. wet vs. dry, respectively, and its effect on ligand transport. All simulations frames (time steps) were analyzed with Mole 2.0,[23] and an ensemble of one hundred frames with an open channel configuration were selected. The output of Mole 2.0 describes the enzyme channel in terms of geometric (radius) and physicochemical (hydrophobicity index – H_I) parameters. Details of the output parameters can be found in the methods section. For each frame, we obtained the channel radius and hydrophobicity index profile at 0.2Å increments. The discrete values for both properties were interpolated and smoothed by a piecewise polynomial fitting spline (Figure 5.4).

To understand the role water hydrogen bond (HB) networking plays inside the channel, we calculated the average ensemble radius for the wet and dry cases and found that the root-mean-squared-variation between the two states was less than

1Å. This shows that, unlike some cytochrome P450s,[33] the water molecules inside the NDO channel do not induce any major conformational changes on the channel architecture that may facilitate the transport of ligands. Instead, the displacement of these water molecules by the ligand from inside the channel to the bulk solvent phase may be the principal source of free energy controlling the transport into the active site. We also calculated the ensemble average H_I index for the wet channel and observed that the solvent exposed entrance region was less hydrophobic than the barrel-shaped mid-section of the channel (Figure 5.4 *bottom*) as expected.

To test the validity of our ligand transport trajectory model, the splines for every frame were used to map each channel onto the sequence of building blocks. In the validation study a set of 45 ligands that have been previously tested experimentally,[7] and computationally[119] in an all-atom model for catalytic activity in NDO were used. Figure 5.5 shows a sample trajectory analysis for six different compounds in a single channel snapshot. For a molecule to have a non-zero probability of being catalyzed by NDO, it must first reach the active site.[119,53] In our model, this corresponds to the ligand having a value of $\Delta G^* < 0$ at all locations along the channel. Figure 5.5 shows that naphthalene, isochroman, and diphenyl sulfide successfully completed transport through the full path to the active site of the enzyme, staying below the cutoff value at all times, and therefore were categorized as "possible substrates" of NDO, matching the experimentally observed results.[7] The detailed mechanism of inhibition by 1H-indole-3-acetate is still unknown, however, it has been proposed that the carboxylic group coordinates to the iron center;[96] meaning that this molecule must also reach the active site in order to inhibit NDO. The results presented in

Figure 5.5 show that the inhibitor was also able to reach the active site of the enzyme, matching the experimental observation. Finally, in the case of fluoranthene and 9,10-dihydro-9,10-methanoanthracene it was observed that only some portions of the trajectory were below the cutoff value, therefore they were correctly categorized as "unlikely substrates."

The trajectory analysis was repeated for all 100 frames to determine if the test ligands reached the active site. In Tables 5.2 and 5.3 we report the probability of entrance into the active site for every ligand tested. We identified a very distinct pattern where the experimentally verified substrates of NDO were successfully transported to the active site through the building block model in more than 90% of the analyzed frames. On the other hand, the experimentally verified poor substrates only completed successful trajectories in less than 10% of the analyzed frames. Overall, our prediction accuracy was 90%, the positive prediction value was 90%, and the negative prediction value was 92% (Figure 5.6 shows the prediction success rate at different discretization intervals). These values are slightly lower than the ones we observed in our previous computational studies.[119] However, the major benefit of this new method is the very fast channel transport prediction time of $\sim$ 1ms/ligand. This is a reduction in computation time of up to 6 orders of magnitude compared to our previously developed all-atom method;[119]

For each frame, we calculated the channel radius and hydrophobicity index (HI) profile using the conversion factor presented in Figure 5.3. The discrete values, for both properties, were interpolated and smoothed by a piecewise polynomial fitting

spline (Figure 5.4). To understand the role water hydrogen bond (HB) networking plays inside the channel, we calculated the average ensemble radius for the wet and dry cases and found that the root-mean-squared-variation between the two was less than 1Å. This shows that the water solvation inside the channel is not inducing any major conformational change on the channel architecture. Instead, this suggests that displacement of water molecules from the channel by the ligand might be a principal source of free energy controlling transport into the active site. We also calculated the ensemble average H_I index for the wet channel and observed that the solvent exposed entrance region was less hydrophobic than the barrel-shaped mid-section of the channel (Figure 5.4 *bottom*), as it would be expected. Table 5.4 shows the prediction times for different currently available methods.

The improvement in the new method in computation time is the result of not having to perform the calculation-intensive all-atom MD simulations for every ligand. Instead, the new method utilizes the pre-calculated non-dimensional free energies (ΔG^*) needed to determine if the transport of a ligand along the set of interlocking building blocks will be favorable or unfavorable. These results show that the new method can be applied as a rapid pre-screening tool before any detailed, yet computationally expensive, all-atom methods is utilized. The simplicity of the mapping procedure also allows the extension of this method to other fields, such as to analyze drug metabolism mediated by the network of hydrophobic channels found in some cytochrome P450s enzymes.[192] Overall, the approach presented here appears to be robust, transferable to other hydrophobic enzyme channels, and capable to elucidate the major geometric and energetic barriers that ligands experience as they

move towards buried active sites. We expect that this method will be a valuable tool for the rational prediction of novel substrates for the production of biofuels, food and agricultural additives, and pharmaceuticals.

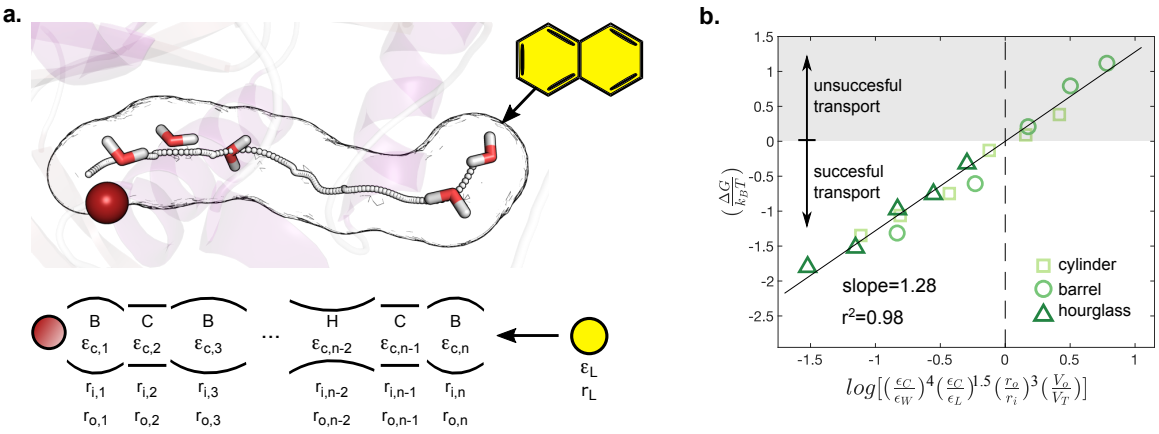

Figure 5.1: Discretization of an enzyme nanochannel for the construction and mapping of the building block model. **a.** (top) Cartoon representation of naphthalene 1,2-dioxygenase (NDO) showing the surface of the channel wall (black), centerline of the channel (white dots), the mononuclear iron at the active site (red sphere), selected water molecules solvating the inside of the channel, and naphthalene (yellow), as the representative ligand. (bottom) Cartoon representing discretization of the NDO channel into building blocks. Each building block shows a schematic of the possible coarse-grained geometries, based on r_i and r_o, and the nonbonded interaction strength (ε) describing the level of wall hydrophobicity (see Figure 5.2 for details). The ligand of interest (yellow circle) is represented by a spherical molecule of uniform hydrophobicity. **b.** Non-linear regression analysis relating dimensionless free energy to characteristic hydrophobicity, geometry, and excluded volume of the building block/ligand combination. The gray region shows the building block geometry and nonbonded interactions for which ligands did not successfully get transported through the building block; thus resulting in an unsuccessful transport.

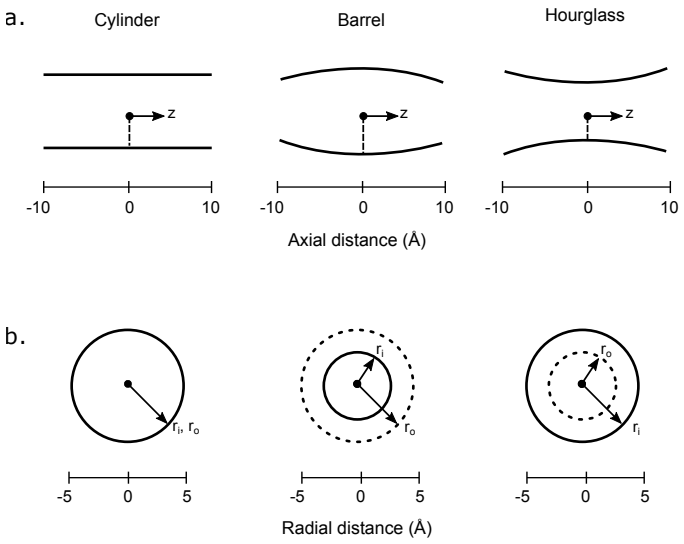

Figure 5.2: Coordinate projection of the system showing the three type of building blocks used in this study. **a.** Projection along the axial coordinate (z-axis), and **b.** Projection along the radial coordinate showing the cylinder, barrel and hourglass geometries. The dashed lines correspond to the radius (r_i) of the channel at z = 0 and the solid lines mark the radius (r_o) at z = $\pm$10Å.

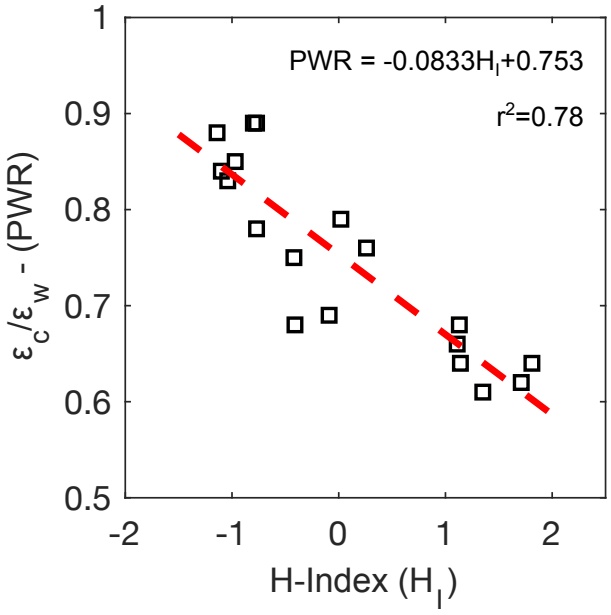

Figure 5.3: Conversion factor between hydrophobicity index (H_I) calculated by MOLE 2.0 and potential well ratio used in our building block model. The symbols show the values for each amino acid, and the dashed red line shows the linear regression.

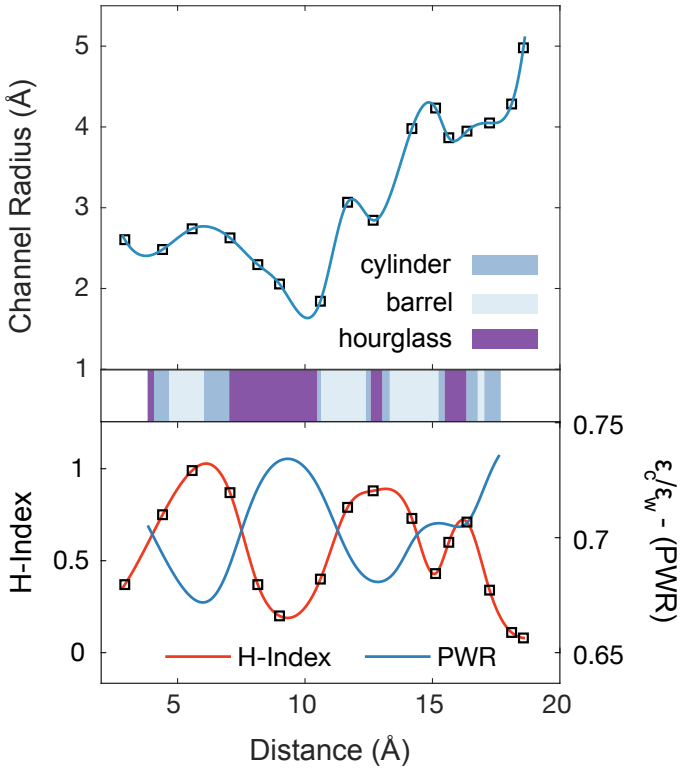

Figure 5.4: Geometric and hydrophobic profile of a single snapshot of naphthalene 1,2-dioxygenase. *(top)* Radius profile of the NDO channel from a single snapshot, symbols correspond to the calculated radius at each step k, and the blue line shows the cubic spline fitting used for the analysis. *(middle)* Heat map showing the type of building block used to model each region along all n steps of the channel, based on the cubic spline. *(bottom)* Hydrophobicity profile of the NDO channel, symbols correspond to the calculated H-Index at each step, k, while the orange line shows the cubic spline fitting, and the blue line shows the conversion of the cubic spline from HI to potential well ratio. Note the two y-axis scales used.

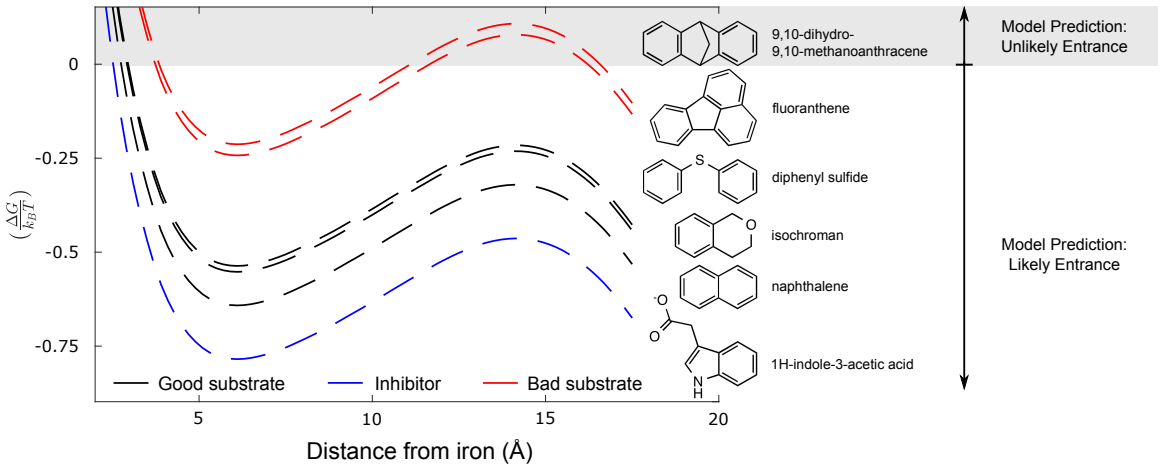

Figure 5.5: Application of the building block trajectory analysis for six different compounds, as shown for a single channel snapshot. Three known substrates (black lines) and one inhibitor (blue line) have successfully transported along the building block model and reached the active site region ($d < 6$Å), thus predicted to be likely substrates. Two poor substrates (red lines) have unfavorable trajectories at the bottleneck region ($d = 12 - 17$Å), thus predicted to be unlikely substrates.

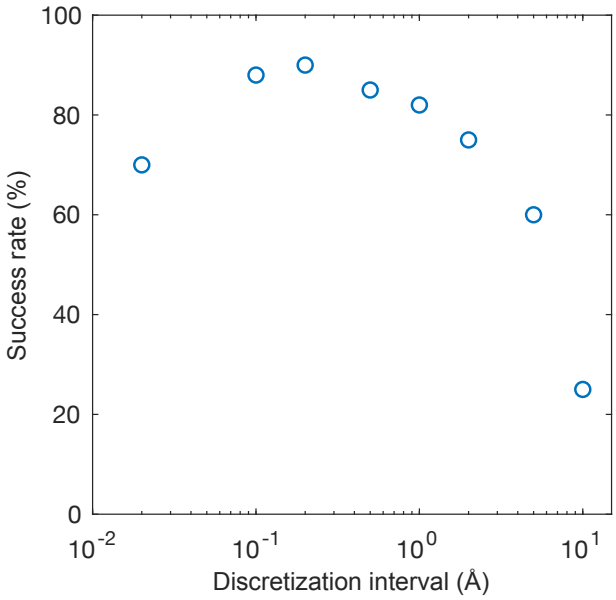

Figure 5.6: Effect of discretization on prediction success. The optimum success rate was reached at a discretization value of 0.2Å.

Table 5.1: Excluded volume fraction (V_O/V_T) used in the nondimensional analysis of the building blocks

PWR	**Cylinder**	**Barrel**	**Hourglass**
0.5	0.544	0.359	0.509
0.6	0.541	0.409	0.522
0.7	0.543	0.425	0.524
0.8	0.541	0.431	0.521
0.9	0.534	0.430	0.520
1.0	0.538	0.428	0.516

Table 5.2: Comparison of the experimental and BB trajectory analysis results for 45 substrates and poor substrates of NDO. [a] Experimental results are from Aukema et. al.[7]

Name	**% Removal[a]**	**% Success**
1,1-diphenylethylene	82	92
hexahydro-s-indacene	35	15
9,10-ethanoanthracene	0	1
9,10-methanoanthracene	0	2
(2S)-flavanone	0	1
(2R)-flavanone	0	0
dihydroacenaphthyelene	100	95
acenaphthyelene	100	96
adamantane	0	10
azulene	100	96
benzene	100	90
benzofuan	100	85
benzophenone	41	30
benzothiophene	100	95
bibenzyl	94	93
1,1'-bicyclohexyl	0	6
biphenyl	100	95
biphneylene	100	93
cis-stilbene	0	6
cyclohexylbenzene	90	90
cyclopropylbenzene	52	92
diphenyl ether	90	95
diphenyl sulfide	87	96

Table 5.3: Comparison of the experimental and BB trajectory analysis results for 45 substrates and poor substrates of NDO. [a] Experimental results are from Aukema et. al.[7] [b] Note that indole-3-acetic acid is an inhibitor and therefore not removed by NDO.

Name	**% Removal**	**% Success**
diphenylmethane	36	6
flavone	35	4
fluoranthene	38	2
indane	100	99
indole	100	98
indole-3-acetic acid [b]	n/A	90
isochroman	100	95
isoflavone	35	2
m-terphenyl	46	9
naphthalene	100	99
phenanthrene	100	90
phenylnaphthalene	21	3
pyrene	7	1
quinoline	100	95
spiro[2.4]hepta-4,6-diene	88	96
spiro[cyclo propane-1-1-indene]	70	92
tetrahydroquinoline	100	92
tetralin	100	92
tras-decalin	58	90
trans-stilebene	82	91
triphenylene	14	2
xanthene	95	94

Table 5.4: Comparison of prediction times for different available methods.

Method	**Time (s)**	**Reference**
Steered Molecular Dynamics	$\sim 10^5 - 10^6$	[18]
Random Accelerated Molecular Dynamics	$\sim 10^5 - 10^6$	[18]
Equilibration Molecular Dynamics	∞	[18]
All-atom Monte Carlo Trajectory	$\sim 10^2 - 10^3$	[119]
Building Block Sequence	$\sim 10^-3$	This work

Chapter 6

Bacterial Aromatic Hydrocarbon Oxygenases And Bridged Ring Hydrocarbons: Computational Studies

6.1 Introduction

Hydrocarbon oxidation reactions are extremely important industrially for processing hydrocarbons in natural gas and petroleum into major commodity chemicals. Moreover, the fate of these chemicals in the environment, due to oil spills or industrial waste, is largely dependent on microbially-mediated enzymatic oxidation reactions.[9,193,194]

Given the industrial and environmental significance of hydrocarbon oxidation reactions, there is increasing interest to use advanced computational tools for modeling hydrocarbon oxygenases to predict substrate specificity for the purposes of bioremediation and synthetic biology. The present study compared oxidation of small bicyclic compounds by three well-characterized oxygenase enzymes both computationally and experimentally: the dinuclear iron enzyme, T4MO, and the Rieske oxygenases, TDO and NDO. X-ray crystal structures exist for each of the three enzymes with and without ligand bound in the active site[195,47,50] making these enzymes especially useful for computational prediction of substrates.

One common feature affecting substrate specificity for these two classes of oxygenases revealed by the crystal structures is the existence of an active site entrance channel. This is in contrast to the extensively modeled P450 oxygenases with surface exposed active sites.[54] Unlike the methods for predicting P450 substrates, predictive algorithms for the dinuclear iron and Rieske oxygenases need to account for bottle-

necks to active site entry. Our group recently developed an algorithm that achieved a positive substrate prediction rate of greater than 90% for NDO by modeling the energetics of substrate entry into the enzyme active site.[119] These prediction data are currently being used to populate the RAPID database cataloging known and predicted enzyme substrates (rapid.umn.edu).

The small bridged compounds represent an understudied class of molecules that pose particular challenges to computational prediction methods for oxygenases because of their unique three dimensional structures. Here we present a computational investigation of bicyclic compounds as oxygenase substrates. This work extends the known substrate range of promiscuous aromatic oxygenases to bridged ring compounds and forwards efforts to use advanced computational tools for predicting substrate selectivity and regioselectivity by these enzymes.

6.2 Computational Methods

6.2.1 Receptor Preparation

X-ray crystal structures of three oxygenases i.e., naphthalene dioxyenase (NDO), toluene dioxygenase (TDO), and toluene-4-monooxygenase (T4MO) were obtained from Protein Data Bank (NDO: 1O7G, TDO: 3EN1, and T4MO: 3Q14).[28,47,195] Computational models were prepared based on the respective X-ray crystal structures of

NDO, TDO and T4MO with bound naphthalene, toluene and p-cresol respectively. Models were optimized prior to docking using the Protein Preparation Wizard in Schrödinger Maestro Suite 2014.[80,82,81] During the optimization process, any inconsistencies in the structure such as missing hydrogens, incorrect bond orders, and orientation of the different functional groups of the amino acids were corrected. The prepared protein was then used for Docking.

Active sites predicted in the respective PDB structures were used in the grid generation wizard of Maestro for all three receptor proteins. A rectangular box confining the translations of the mass center of the ligand defined the binding site. It was defined by keeping the co-crystallized ligand at the center of the box in the receptor. For each protein structure, a grid box of default size was centered on the corresponding ligand position. All default parameters were used except H-bond and hydrophobic constraints, which were included during grid generation.

6.2.2 Ligand Preparation

The structures of the three bridged compounds were retrieved from the NCBI PubChem database (Figure 6.1): Norbornane (bicyclo[2.2.1]heptane, CID: 9233]; Norbornylene (bicyclo[2.2.1]hept-2-ene, CID: 10352); and Norbornadiene (bicyclo[2.2.1]-hepta-2,5-diene, CID: 8473). These compounds were prepared prior to docking using the LigPrep application in Schrödinger Maestro Suite 2014.[80,82,81] LigPrep performs optimization of ligand structures by: i) converting structures from two to three dimen-

sions; ii) correcting improper bond distances; iii) determining bond orders; iv) sgenerating ionization states; v) establishing correct chiralities; and vi) and determining structure of minimum energy. The force field OPLS_2005 was used for minimization and the ionization states were generated at the default pH of 7.0 ± 2.5. The ligand structures prepared by LigPrep were then used for Docking.

6.2.3 Docking

In order to know the binding mode of bridged compound with the three oxygenases, molecular docking of NDO, TDO and T4MO models were carried out using XP (extra precision) module in which model was docked in different orientations with the ligands. Performance of docked complexes was analyzed based on the parameters: Glide score, H-bonding score, H-bond interaction and bond distance (Å). For ligands with multiple Glide scores in different orientations, the best scoring pose with XP Glide score was taken.

6.2.4 Estimation of Ligand Binding Energy

The ligand binding energy of each bridged compound to the three oxygenases was estimated using Prime MMGBSA module in Schrödinger Suite 2014.[80,82,81] The total

free energy of binding, δG_{bind} (kcal/mol) was estimated by the software as follows:

$$\begin{aligned}\delta G_{\text{bind}} = &[\text{energy of complex}]_{\text{min}} \\ &- [(\text{energy of ligand})_{\text{min}} + (\text{energy of receptor})_{\text{min}}]\end{aligned} \tag{6.1}$$

The best poses of the docked complexes were then chosen to obtain the binding free energy calculation. Prime MM-GBSA is a method that combines Optimized Potential for Liquid Simulations-All Atoms (OPLS-AA) force field, molecular mechanics energies (EMM), an Generalized Born (GB) solvation model for polar solvation, and a non-polar solvation term composed of the non-polar solvent accessible surface area and van der Waals interactions. We then used this score to rank the substrate-protein docked complex.

Glide docking and Prime/Molecular Mechanics Generalized Born Surface Area (Prime/MM-GBSA) calculations were applied to predict the binding mode and free energy for the bridged substrate series for each oxygenase. Docking results were obtained using Glide docking protocol. Then, the Prime/MM–GBSA method based on the docking complex was used to predict the binding-free energy. The obtained ligand poses were minimized using the local optimization feature in Prime, whereas the energies of the complex were calculated with the OPLS_2005 force field and Generalized Born/Surface Area continuum solvent model. During the simulation process, the ligand strain energy was also considered.

6.3 Results and Discussion

6.3.1 General properties of the oxygenase enzymes and bridged ring compounds

The natural substrates and oxygenation reactions of the naphthalene dixoygenase (NDO), toluene dioxygenase (TDO) and toluene-4-monooxygenase (T4MO) are shown in Figure 6.1A. While these enzymes each have a broad substrate range, the size of the active site imparts a degree of substrate discrimination. Substrates larger than the active site are excluded. Based on the volume comparison of the bridged compounds to the natural substrates, toluene and naphthalene, one predicts the bridged compounds should not be excluded from the active site based on size (Figure 6.1B). Furthermore, comparison of compound volumes to the active site volumes, calculated by CASTp server using 1.4Å probe radius[196] supports the same prediction. Average active site volume for NDO, TDO and T4MO are $700,300$ and 290 $Å^3$ respectively. Based on volume calculations alone, one predicts these compounds to be potential substrates of each of the three oxygenases. However, despite the large active site volume of NDO in comparison to the volume of the three bridged ring compounds (Figure 6.1), norbornane, nobornylene and norbornadiene were reported not to be substrates for NDO.[7] Therefore, additional factors prevent these bridged compounds from being oxygenated by NDO.

6.3.2 Analysis of active site binding energy and accessibility

Two possible explanations for the lack of reactivity of NDO with the bridged compounds are poor energetics of substrate binding and the inability of the molecules to traverse the channel leading into the enzyme active site. First, to investigate active site docking computationally, glide docking and Prime/Molecular Mechanics Generalized Born Surface Area (Prime/MM-GBSA) calculation were applied to predict the binding mode and free energy for the bridged substrate series and several control compounds in each oxygenase (Table 6.1). The calculated free energies of binding of the known substrates naphthalene and toluene for NDO and TDO, respectively, are similar, while the calculated binding for T4MO is somewhat poorer (less negative). Benzene serves as a lower boundary as it is a smaller molecule and it is known to be a poorer substrate for all three enzymes than their respective substrates. The binding energies for three bridged compounds were intermediate between the positive and negative controls for NDO and T4MO but were lower (better binding) for TDO. These data suggests that these bridged compounds may be substrates for TDO, but do not preclude the compounds from being substrates for NDO and T4MO. Therefore, an additional explanation for the lack of reactivity in NDO was sought.

We considered the possibility that unlike the planar natural substrate, naphthalene, the three dimensional nature of the bridged compounds prevents entry into the active site of NDO. Entry of the compounds into the active site was analyzed computationally using a recently developed method that measures non-bonded inter-

actions (electrostatic and Van der Waals) of a compound as it traverses the channel to the active site in a representative 10% of the duration of a 20ns molecular dynamics simulation.[119] Computationally, the bicyclic compounds were able to access the active site of NDO with a frequency greater than the native substrate, naphthalene. Therefore, active site accessibility cannot explain the lack of reactivity of NDO with the bridged compounds.

6.3.3 Substrate orientation and product prediction

We then hypothesized that the bridged compounds may not bind in the NDO active site in a productive orientation or position. With NDO, benzene positioning in the distal pocket of the enzyme more than 5Å away from the active site iron is thought to explain why it is a poor substrate, as observed experimentally by Lee.[197] In light of this, we examined the preferred docking orientations and positions of norbornane in each of the active sites compared with the orientation and position of known substrates solved in X-ray crystal structures of the respective enzymes with bound ligands (Figure 6.2). The large active site of NDO with respect to TDO and T4MO is represented with the 3-ring aromatic hydrocarbon anthracene bound in cyan. The TDO and T4MO active sites have toluene and p-cresol bound, respectively, in cyan. There is presently no available structure for T4MO with the substrate, toluene, bound so comparisons were made with the product bound structure. Comparison of norbornane positioning with the ligands bound in each enzyme active site shows that the bridged substrates are not ideally positioned in any of the oxygenases, however, their

proximity to the iron and the activated oxygen species is closer with TDO and T4MO than with NDO. Relevant distances are reported in the supplemental material.

For each X-ray crystal structure with ligand bound, the obtained ligand poses were minimized using the local optimization feature in Prime. The dioxygen bound to the iron was added to the model based on the positioning of oxygen in the ligand-bound structure PDB: 1O7N.[28] Figure 6.2a shows that this model predicts that anthracene is attacked from the bottom face of the A ring to produce the specific diastereomer cis-1,2-dihydroxy-1,2-dihydroanthracene. This is the only product observed experimentally for anthracene with an enantiomeric excess over >98%.[25]. The computationally determined positioning of nonbornane coincides with the B and C rings of anthracene, the two rings that are not oxidized by NDO to any detectable degree, despite the 9,10 positions (B ring) of anthracene being more reactive with chemical oxidants.[198] This suggests that although the bridged substrates can access the active site, they do not bind in a suitable position for oxidation by NDO.

Compared to the predicted binding position in NDO, norbornane binds closer to the activated oxygen atoms of the TDO and T4MO active sites (Figures 6.2b and 6.2c). If oxygenation of a bridged compound were to occur, products can be predicted based on the docking positions shown in Figure 6.2. The regiospecificity of oxygenation by TDO and T4MO is predicted to favor hydroxylation at C2 or C3 to the side of the bridged linkage in preference to C1 and C7, the bridgehead and bridge carbons, respectively. With TDO, cis-dihydrodiols are the expected products for the double bond-containing compounds, norbornene and norbornadiene, based

on docking orientation and chemical selectivity. The enantiospecificity of the TDO products is difficult to predict based on the docking position alone. The C2 and C3 endo and exo hydrogen to iron distances are nearly the same. Exo products are predicted to be somewhat favored because the predicted O_2 channel opens on the exo side of the bridged compound. Furthermore, the exo products are more energetically favored via chemical reactivity.[199]

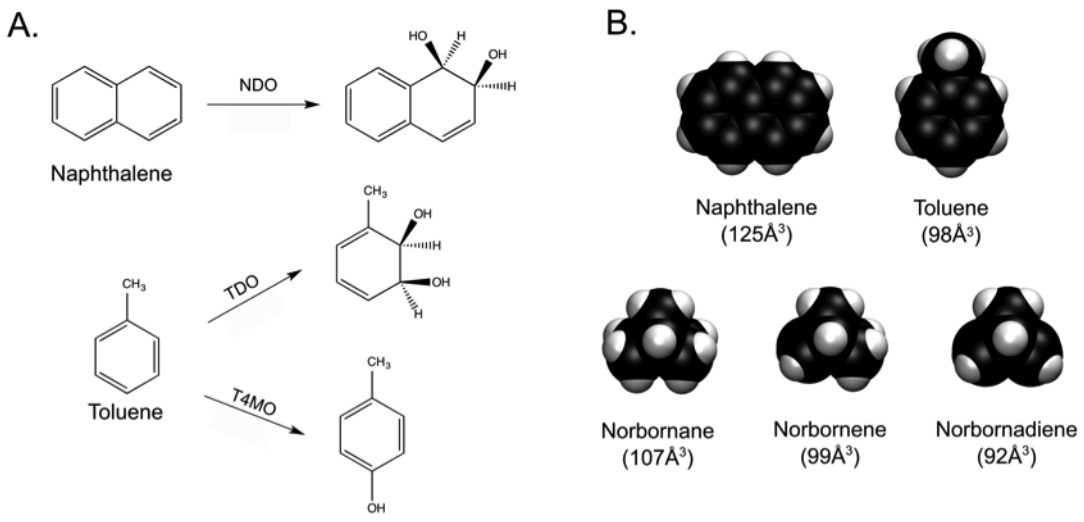

Figure 6.1: Aromatic hydrocarbon oxygenases and the compounds studied here (A) Native reactions of naphthalene dioxygenase (NDO), toluene dioxygenase (TDO), and toluene 4-monooxygenase (T4MO). (B) Space filling model of native substrates and bridged compounds with calculated molecular volumes.

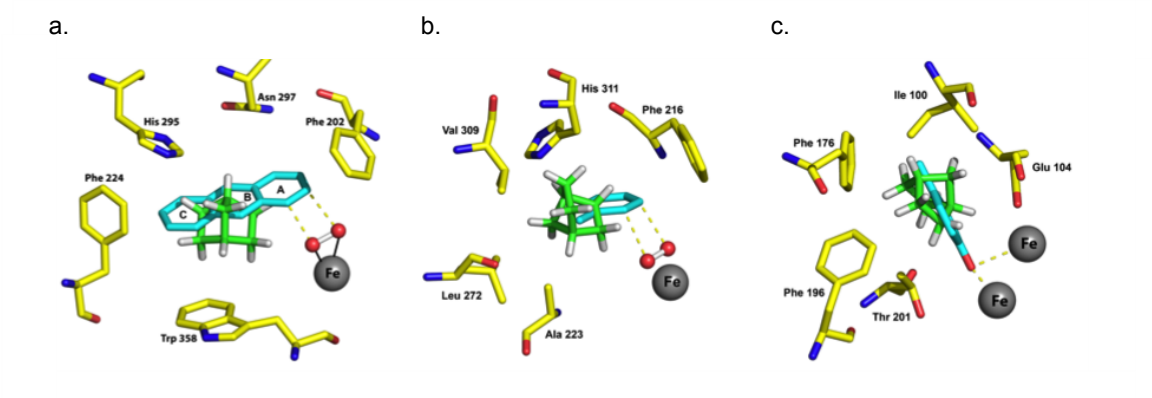

Figure 6.2: Representation of NDO, TDO and T4MO active sites with norbornane docked structures. a) Anthracene bound NDO active site (PDB: 2HMM), b) Toluene bound TDO active site (PDB: 3EN1) and c) p-cresol bound T4MO active site (PDB: 3Q14). Dioxygen molecules are shown as ball and stick (in a and b). Norbornane molecules are shown in green stick with white hydrogen. NDO substrate anthracene, TDO substrate toluene and T4MO product pcresol are modeled as stick in cyan.

Table 6.1: Free energy calculations for binding of bridged and aromatic ring compounds to each of the three oxygenases. Energies are listed in kcal/mol. The preferred substrate correspond to naphthalene or toluene, respectively.

	NDO	TDO	T4MO
Preferred Substrate	-39.2	-40.3	-28.8
Benzene	-24.8	-24.0	-17.5
Norbornene	-36.8	-46.6	-25.2
Norbornadiene	-36.6	-46.0	-23.9
Norbornane	-36.6	-46.0	-23.9

Chapter 7

Reaction Activity Prediction Identification

The Rapid website can be access at: `rapid.umn.edu`

7.1 Predicting Microbial Biocatalysis and Biodegradation

Models are primarily developed to predict the behavior of systems whose response cannot be measured experimentally. The main impediments for measuring the responses can be due to the scale of the system of interest, or in other cases, the measurements would greatly disrupt the natural environment of the system yielding non-representative data. In predicting biocatalysis and biodegradation the limitation is imposed by sheer size of the analyzable population.

7.1.1 Motivation for predicting Microbial Biocatalysis and Biodegradation

According the American Chemical Society there are approximately 10 million known organic compounds. Despite all the developments in experimental microbiology and rapid high-throughput screening, the number of known biotransformations and their responsible enzymes only covers approximately 0.01% of the known organic universe. Most of this information is stored and curated in online databases, Section 7.1.2 gives a brief description of some of these databases. The ability of accurately predict biocatalysis and biodegration has enormous implications[6], and the true value of such capability will be reflected by the fact that:

- Amount of experimental work required to find lead compounds for biocatalysis can be reduced by orders of magnitude.

- Rational design and engineering of metabolic pathways will be a possibility.

- Networks and consortia of microorganisms can be designed to achieve non-natural full biodegradation of xenobiotics.

- Libraries containing biosynthesized fragments with large number of chiral centers can be built and used by the pharmaceutical industry.

- Many lead compounds generated by rational drug design will be able to synthesized with high enantiomeric purity (a requirement by the Federal Drug Administration).

- Exposure to hazardous chemicals will be limited to only those chemical compounds likely to exhibit biodegradative capabilities.

- Regulatory agencies and researchers will be able to study the environmental fate of materials.

- Companies can avoid the commercialization of compounds which are likely to be transformed into hazardous chemicals when exposed to certain enzymes.

7.1.2 Current computational tools for predictive purposes

The Internet has become an ideal platform to host and distribute information about biodegradation and biocatalysis in the form of databases and commercial prediction software. Some of the most common computational tools currently available are the following:

- Curated databases containing experimentally verified chemical reactions catalyzed by enzymes (Figure 7.1 names some specific examples)[200]. These databases are a good starting point for studying biotransformations, however, the content is solely based on experimental results thus, the amount of possible reactions is extremely limited. In addition, it is common that enzymatically-catalyzed reactions published in the literature are not updated frequently leaving a gap between the known and readily available information.

- Computational prediction systems based on empirical biotransformation rules capable of outlining possible pathways. The benefit from these pathway prediction systems is that they provide a possible transformation pathway for the chemical. However, the disadvantage is that the biotransformation rules to obtain the pathways are not capable to providing likely organisms to carry out a given reaction in the pathway[201,129,202,203].

- Quantitative structural-activity relationship models (QSAR) are regression mod-

els used to predict the biological activity. This models are based entirely on the chemical structure of the compound being studied. Therefore, they only provide a very general overview of the possibility of being biotransformed, but do not tell anything about what enzymes would be capable of carrying out such reactions. Also, this method requires a training set which inherently biases the prediction capabilities of the method to find results similar to those trained with, limiting the prediction of novel substrates[204].

- Advanced computational methods attempt to generalize the highly intricate and complex thermodynamic interaction between the chemical of interest and the enzyme active site by using semi-empirical force fields. This technique is commonly known as Molecular Docking, and it provides an extremely simplified free energy of binding. The advantage of this method is that it is capable of determining the best position, if any, of the chemical in the active site pocket. However, one of the main disadvantages is that this is only a representation of the enzyme at one instance in time[81,82].

- Molecular Dynamic Simulations provide a better description of the thermodynamic interactions between the chemical and the active site over a period of time. However, the main disadvantage of this technique is that it is extremely computationally expensive, and simulations can only be done for key chemicals, thus this is not a high-throughput technique.

7.1.3 Bridging the gap

It is evident that there is a great potential in developing methods for predicting biocatalysis. Several attempts have been made already, as described in Section 7.1.2, to provide an insight into the fate of a chemical when exposed to enzymes. However, **there is a clear gap between the current computational tools and the desire to elucidate novel chemical reactions in a shorter period of time, or without requiring the a high level of expertise and knowledge on the mechanism of action of specific enzymes**. This means that scientists and engineers have to currently spend large amounts of time studying and understanding the way that certain enzymes work to start generating predictions. This is then followed by laborious wet lab work which includes exposure to, sometimes highly toxic, chemicals. Finally, many of the attempts to find the right enzyme to catalyze the desired reaction fail and the whole process needs to be repeated. Therefore, the overall goal of this chapter is to develop a computational model based on the thermodynamics of nonbonded interactions between Rieske nonheme iron enzymes and chemical compounds, to rapidly generate accurate substrate and product predictions. There is a high up-front cost to develop all the necessary modules of this new computational tool. However, once all the modules are up-and-running they will be able to be repeatedly used with any combination of enzyme and chemical.

7.1.4 RAPID – For identifying and exploiting enzyme catalysis

Motivated by the lack of a reliable tool capable of predicting relevant enzyme catalysis, the overall goal of RAPID is to provide an easily accessible and user-friendly computational tool capable of identifying broad-specificity enzymes likely to catalyze a given reaction. Therefore the search-and-predict online database is accessible to the general public under the name of RAPID (**R**eaction **A**ctivity **P**roduct **ID**entification) available at the URL: `rapid.umn.edu`.

As for any engineering system, it is essential to identify the parameters that define the model, for RAPID these are the inputs and outputs. Figure7.2 shows an overview of the expected inputs into the system and the possible outputs after performing all the calculations in the model.

The RAPID website was designed to properly work on the basis of six possible different Substrate-Product Pairs (S-PP) reactions, as listed below:

- One substrate : One Product $(1:1)$
- One substrate : Many Products $(1:\infty)$
- One substrate : No Products $(1:\neg)$

- Many substrates : One Product ($\infty : 1$)

- Many substrates : Many Products ($\infty : \infty$)

- Many substrates : No Products ($\infty : \neg$)

Furthermore, each of these combinations of S-PP reactions, except for the ones with no products, will also provide the enzyme responsible for its catalysis. And as shown in Figure 7.2, one or many enzymes could be capable of catalyzing the predicted S-PP reaction.

Similar to the chemical universe described in Section 7.1.1, the enzyme universe consists of millions of proteins, and trying to analyze all of these enzymes would be unrealistic. Therefore, the first iteration of the development of RAPID will consist of 5 broad-specificity enzymes (BSE). For the purposes of our new web-database, an enzyme can be considered a BSE if it meets two important criteria: i) catalyze at least 15 different S-PP reactions; and ii) have a resolved crystal structure deposited in the Protein Databank (PDB).

Although not drawn to scale, Figure 7.3 shows that one of the main challenges in deciding the BSE is that the number of crystal structures deposited in the PDB is not proportional to the number of known S-PP reactions. Therefore, this imposes a challenge as a diverse set of enzymes is desired in order to capture and understand as many reactions as possible. Table 7.1 shows the list of the 11 BSS enzymes that

RAPID will use during the first stage of prediction development.

Figure 7.4 shows the databases that were used to initially populate RAPID. The structure of the website and server is outlined in Figure 7.5 and the currently live page is shown in Figure 7.6.

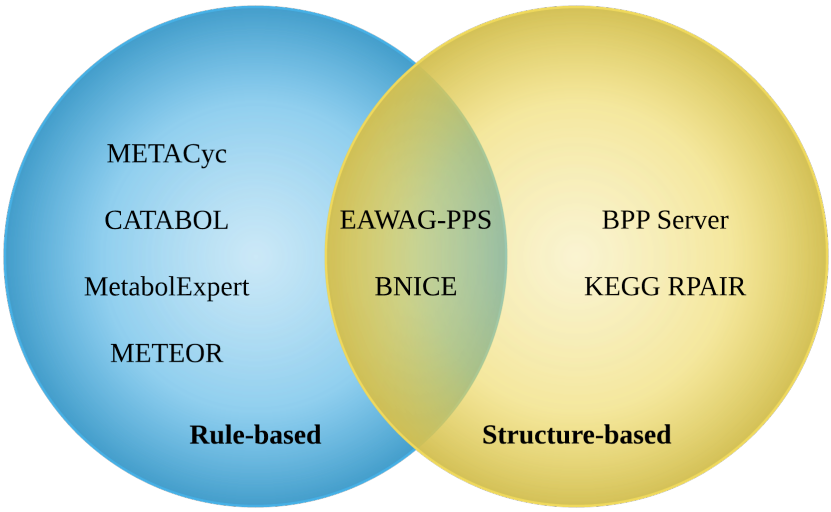

Figure 7.1: Rule and structure based prediction databases available

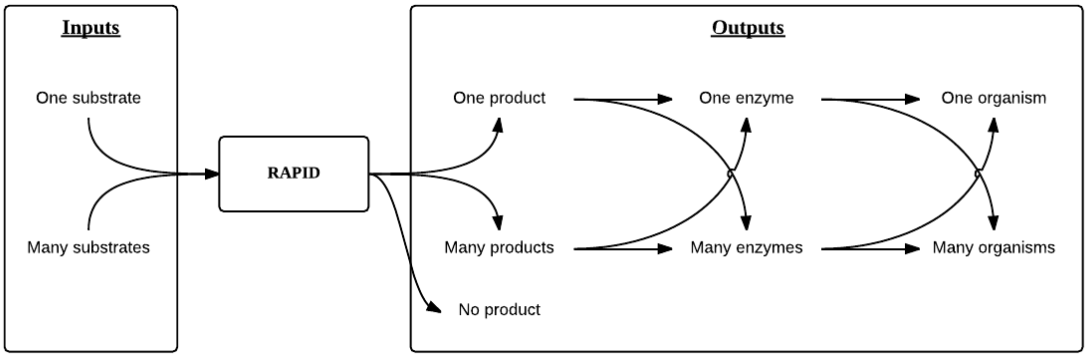

Figure 7.2: General overview of RAPID, and identification of the input parameters required to define the system, followed by the possible outputs.

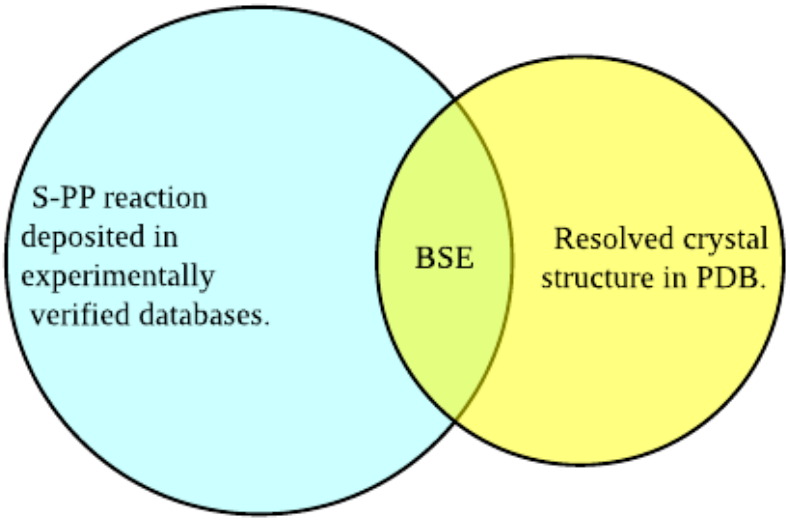

Figure 7.3: Venn diagram showing the intersection of criteria required for an enzyme to be considered a BSE in RAPID.

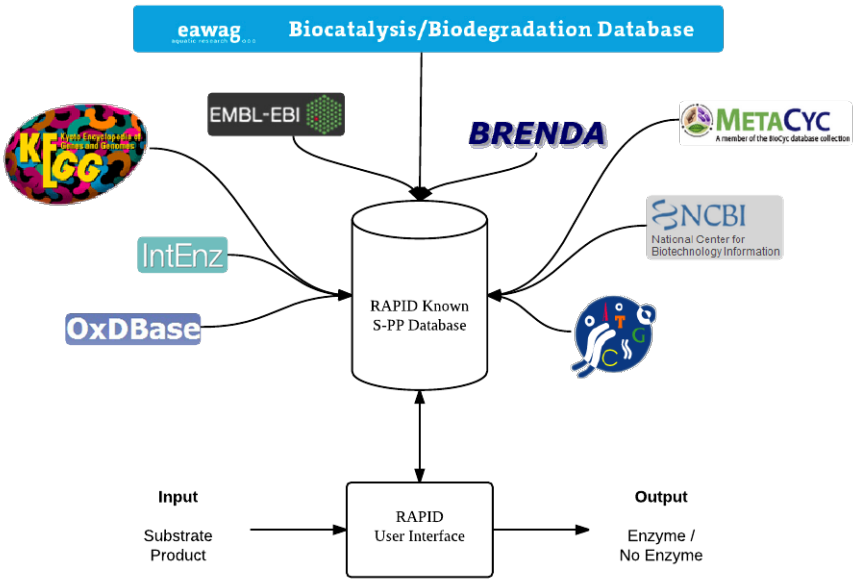

Figure 7.4: Experimentally verified S-PP reactions obtained from public access databases will be stored internally in RAPID. The cached information will be quickly retrievable when an user performs a query.

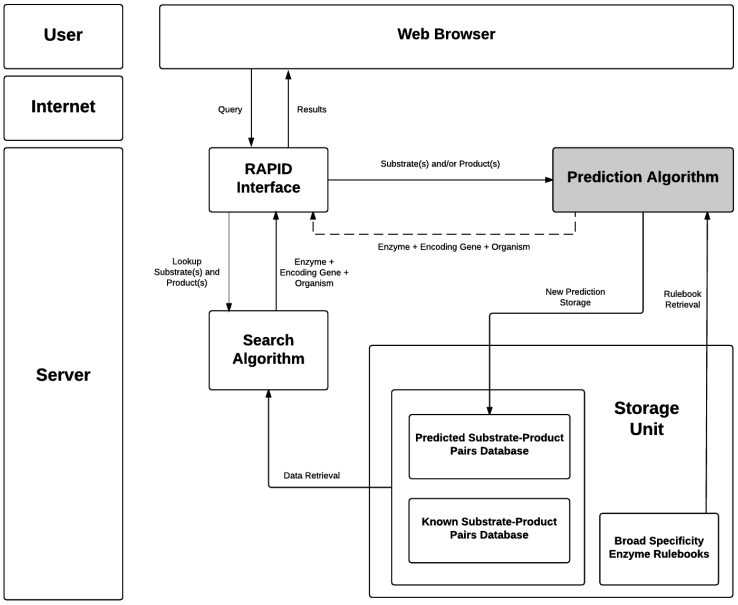

Figure 7.5: Outline of the RAPID server and the integration scheme between the search and prediction algorithms.

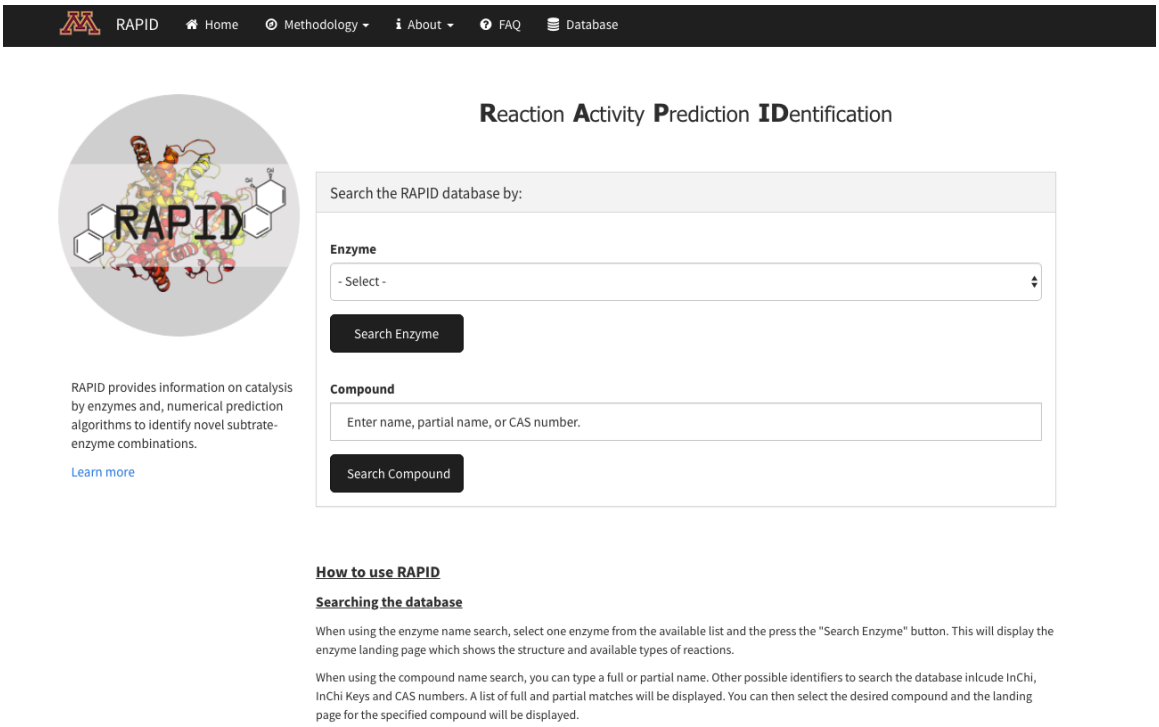

Figure 7.6: Biodegradative and biocatalytic reactions catalyzed by Naphthalene 1,2-dioxygenase from *Pseudomonas sp.* NCIB-9816. Figure adapted from Wackett et. al.[6]

Table 7.1: List of Broad-substrate specificity Enzymes used by RAPID and the organisms that express them.

Enzyme	Organism
Naphthalene 1,2-dioxygenase	*Pseudomonas sp.* NCIB 9816-4
Toluene 2,3-dioxygenase	*Pseudomonas putida* F1
Biphenyl 2,3-dioxygenase	*Sphingobium yanoikuyae* B1
Biphenyl 2,3-dioxygenase	*Paraburkholderia xenovorans* LB400
(S)-hydroxynitrile lyase	*Hevea brasiliensis*
Cytochrome P450(BM-3)	*Bacillus megaterium* ATCC 14581
D-hydantoinase	*Geobacillus stearothermophilus* SD1
Haloalkane dehalogenase	*Sphingomonas paucimobilis* UT26
Haloalkane dehalogenase	*Xanthobacter autotrophicus* GJ10
Toluene 4-monooxygenase	*Pseudomonas mendocina* KR1
Oxo lactonase SsoPox	*Sulfolobus solfataricus* P2

Chapter 8

Research Summary and Future Work

8.1 Research Summary

In this dissertation, studies were undertaken to develop computational models capable of modeling the biotransport of small organic molecules into the active site of broad-substrate specificity enzymes. Two different types of models were developed, tested, and validated: i) an all-atom method; and ii) a non-dimensional generalized method. The computational methods presented in this dissertation will be a valuable tool for the rational prediction of novel substrates for the production of biofuels, food and agricultural additives, and pharmaceuticals precursors. The important results and conclusions drawn from the studies used to developed these methods are summarized as follows:

CHAPTER 2: Simulation of the bottleneck controlling access into a Rieske active site: predicting substrates of naphthalene 1, 2-dioxygenase

In this chapter, an all-atom Monte Carlo simulation algorithm was developed to model the transport of ligands into the active site of Naphthalene 1,2-dioxygenase (NDO). NDO was used as the model enzyme to represent the Rieske non-heme iron dioxygenase family. The model was validated against a set of experimentally verified substrates and poor substrates, yielding 92% accuracy. The main assumptions for the development of the model were: i) on the basis of crystallographic evidence, any induced-fitting changes in the enzyme as the ligand binds in the active site were; ii) NDO is stiff and that all of the major conformational changes were observed

within a medium-length MD simulation; iii) the active-site cavity of NDO is highly hydrophobic, allowing us to ignore solvation effects; and iv) vibrational effects of the ligand were ignored only roto-translational effects were considered. All these assumptions would need to be revised if the algorithm were to be implemented for to a different BSS enzyme family. The pseudocode for the implementation of the algorithm is presented in Appendix C.

CHAPTER 3: *In silico* Identification of Bioremediation Potential: Carbamazepine and Other Recalcitrant Personal Care Products

In this chapter, the all-atom algorithm developed in Chapter 2 was used to identify possible enzymes that can biodegrade recalcitrant and emerging pollutants such as prescription drugs and personal care products (PCPs). Based on the all-atom model it was predicted that carbamazepine, an anti-epileptic drug, can be biodegraded by the enzyme biphenyl 2,3-dioxygenases, expressed in *Paraburkholderia xenovorans* B1. This prediction was experimentally tested and verified. Three other Rieske enzymes were also tested and correctly predicted to not be viable options for the biodegradation of this pollutant. A followup set of 22 PCPs were computationally modeled to predict if any enzyme would be capable of biodegrading them. The computational results suggested that various members of the Rieske family of enzymes would be possible candidates to biodegrade these water pollutants. Experimental results verified that 12 of these 22 PCPs were correctly predicted to be biodegraded by Rieske enzymes. Remarkably, some of these compounds have not previously been demonstrated to be biodegradable by a single bacterial strain.

CHAPTER 4: Role of Water Hydrogen Bonding on Transport of Small Molecules Inside Hydrophobic Channels

In this chpater, a systematic analysis of water networking inside non-cylindrical hydrophobic structures in order to elucidate the role water hydrogen bonding plays on ligand transport was performed. It was established that hydrophobic confinement promotes formation of a water exclusion zone adjacent to the channel walls, and this effect is influenced by the geometry such that a concave channel amplifies the hydrophobic exclusion effect as compared to cylindrical or convex geometries of the same non-bonded interaction strength. Within the water exclusion zone, frictional resistance is reduced, significantly accelerating the hydrophobic ligand transport across. Accelerated hydrophobic ligand transport through hydrophobic geometries is also explained from a thermodynamic perspective of disrupting water hydrogen bonding network. At one extreme, we found that the entropic contributions to the structural order of water inside the geometry might preclude the ligand from entering, as this would cause a disruption on the fragile hydrogen bond network of the confined and energetically "frustrated" water molecules. On the other extreme, if the enthalpic contributions are very high, the water molecules are displaced to the bulk as the small hydrophobic ligand adsorbs onto the hydrophobic wall, slowing down diffusion.

CHAPTER 5: Prediction of Ligand Transport Along Hydrophobic Enzyme Nanochannels

In this chapter, a coarse-grained non-dimensional model for prediction of lig-

and transport inside hydrophobic enzyme nanochannels was developed. To reduce the excessive computational requirement for calculating all pairwise interaction potentials, a simple discretization (slicing) procedure with which a hydrophobic channel inside an enzyme is represented as a sequence of building blocks was performed. Each building block is defined by three parameters to describe its geometry and physicochemical characteristics: i) entrance radius: r_i; ii) midpoint radius: r_o; and iii) the intermolecular nonbonded interaction strength (ε). The nonbonded interaction strength of the building block, ε_C, was defined in terms of the Lennard-Jones potential. Similarly, the ligand was modeled as a sphere of uniform hydrophobicity represented by the nonbonded interaction strength, ε_L. The building block geometric parameters were non-dimensionalized as follows: r_o/r_i. The nonbonded strengths of the building block, and the ligand were non-dimnesionalized with respect to the potential well of a SPC water molecule ($\varepsilon_C/\varepsilon_W$, and $\varepsilon_L/\varepsilon_W$, respectively). The overall prediction accuracy of this model was 90%, the positive prediction value was 90%, and the negative prediction value was 92%. The major benefit of this new method is the very fast channel transport prediction time of $\approx$ 1ms/ligand. This is a reduction in computation time of up to 6 orders of magnitude compared to all-atom methods. This non-dimensional method is robust, transferable to other hydrophobic enzyme channels, and capable to elucidate the major geometric and energetic barriers that ligands experience as they move towards buried active sites. This tool will be a valuable for the rational prediction of novel substrates for the production of biofuels, food and agricultural additives, and pharmaceuticals.

CHAPTER 6: Bacterial Aromatic Hydrocarbon Oxygenases And Bridged Ring

Hydrocarbons: Computational Studies

In this chapter, the thermodynamic nonbonded interactions between ligands and the active site of three enzymes was studied: naphthalene 1,2-dioxygenase (NDO), toluene 2,3-dioxygenase (TDO), and toluene 4-monooxygenase (T4MO). Three cyclic bridge compounds were studied, instead of the the conventional planar aromatic hydrocarbons substrates of Rieske enzymes. The active site of NDO, TDO and T4MO were found be different in size and have varying physicochemical properties. The three cyclic compounds were able to traverse the access tunnels of all three enzymes. However, in NDO, as soon as the cyclic compounds entered the active site they were 'stuck' and could not reach the distal end were the reactive iron is located. On the other hand, in TDO and T4MO, the cyclic compounds were able to position closer to iron in order for them to be reacted upon. The prediction that NDO would not be capable of biodegrading these cyclic compounds was later confirmed by experiments.

Chapter 7: Reaction Activity Prediction Identification

In this chapter, the need for an online repository of reactions by broad-substrate specificity enzymes is outlined. The capabilities of the Rapid website are described (`rapid.umn.edu`) and the schema of data processing is provided. Although Rapid is in its initial stages of development it has already been populated with 11 BSE enzymes and has over 1200 known reactions and over 50 predicted reactions.

8.2 Future Work

There are three main areas in which the work presented in this dissertation can be expanded:

i) incorporation of neglected terms into the all-atom model (Section 8.2.1).

ii) expansions to the building blocks model (Section 8.2.2).

iii) integration of prediction algorithms into the RAPID website (Section 8.2.3).

8.2.1 Future Development of All-atom Algorithm

For the all-atom model there were four main assumptions that can be addressed in the future: i) on the basis of crystallographic evidence, I ignored any induced-fitting changes in the enzyme as the ligand binds in the active site; ii) I assumed that NDO is stiff and that all of the major conformational changes were observed within a medium-length MD simulation; iii) I assumed that the active-site cavity of NDO is highly hydrophobic, allowing us to ignore solvation effects; and iv) I ignored any vibrational effects of the ligand and considered only roto-translational effects. By incorporating these effects, the number of enzyme channels that can be studied would increase as more cases would be able to be tested.

8.2.2 Future Development of Building Blocks

I propose three possible expansion routes for the building block model. First, altering the PWR of the ligand would affect its transport properties. A lower PWR would displace more water molecules from the BB interior. The increase in water exclusion zone size could potentially further increase D_L. Second, ligands of non-spherical shapes can be considered. For instance, an ellipsoid would be more suitable to study the transport properties of large planar compounds. Alternatively, two bonded spheres to represent a single ligand would allow us to model substrates with asymmetrical hydrophobicity. Third, incorporating hydrophilic effects. The presence of a polar group in BB walls, or on the ligand, would considerably change the way that surrounding water molecules behave. All of these modifications would affect the transport properties of ligands and allow expanding the scope of the building blocks.

8.2.3 Future Development of Rapid Website

The major improvement and perhaps most pressing future work is the incorporation of the prediction methods into the Rapid website. The first stage of work would include the prediction of substrates of the NDO enzyme in a streamlined online manner. This would allow users of the website rapidly determine if untested compounds are likely to be catalyzed by NDO. The next step would be to develop the building blocks for other enzymes such as BPDO and TDO. This would allow for the expansion of the

predictive capabilities of the website. In terms of integrating the all-atom model, it would be beneficial for the user to know possible protein engineering sites. These predictions would inform the user of what amino acid mutations are the most likely to change the activity of the enzyme by allowing the entrance of substrates into the active site.

Bibliography

1. Dagley, S. Lessons from biodegradation. *Annu Rev Microbiol* **41,** 1–23 (1987).
2. Singh, R., Kumar, M., Mittal, A. & Mehta, P. K. Microbial enzymes: industrial progress in 21st century. *3 Biotech* **6,** 174 (Dec. 2016).
3. Li, S., Yang, X., Yang, S., Zhu, M. & Wang, X. Technology prospecting on enzymes: application, marketing and engineering. *Comput Struct Biotechnol J* **2,** e201209017 (2012).
4. Choi, J. M., Han, S. S. & Kim, H. S. Industrial applications of enzyme biocatalysis: Current status and future aspects. *Biotechnol Adv* **33,** 1443–54 (Nov. 2015).
5. Staff, B. R. *Enzymes in industrial applications, Global markets* tech. rep. (2018).
6. Wackett, L. P. & Hershberger, C. D. Biocatalysis and biodegradation (2001).
7. Aukema, K. G., Kasinkas, L., Aksan, A. & Wackett, L. P. Use of silica encapsulated Pseudomonas sp. strain NCIB 9816-4 in biodegradation of novel hydrocarbon ring structures found in hydraulic fracturing waters. *Appl Environ Microbiol* **80,** 4968–4976 (Aug. 2014).
8. Ha, H., Mahanty, B., Yoon, S. & Kim, C. G. Degradation of the long-resistant pharmaceutical compounds carbamazepine and diatrizoate using mixed microbial culture. *Journal of Environmental Science and Health Part a-Toxic/Hazardous Substances & Environmental Engineering* **51,** 467–471 (May 2016).
9. Leahy, J. G. & Colwell, R. R. Microbial degradation of hydrocarbons in the environment. *Microbiol Rev* **54,** 305–15 (Sept. 1990).
10. Li, Z., Sobek, A. & Radke, M. Fate of Pharmaceuticals and Their Transformation Products in Four Small European Rivers Receiving Treated Wastewater. *Environmental Science & Technology* **50,** 5614–5621 (June 2016).
11. Payne, R. B., Fagervold, S. K., May, H. D. & Sowers, K. R. Remediation of Polychlorinated Biphenyl Impacted Sediment by Concurrent Bioaugmentation with Anaerobic Halorespiring and Aerobic Degrading Bacteria. *Environmental Science & Technology* **47,** 3807–3815 (Apr. 2013).

12. Seo, J. S., Keum, Y. S. & Li, Q. X. Bacterial Degradation of Aromatic Compounds. *International Journal of Environmental Research and Public Health* **6,** 278–309 (Jan. 2009).
13. Fishelovitch, D., Shaik, S., Wolfson, H. J. & Nussinov, R. Theoretical characterization of substrate access/exit channels in the human cytochrome P450 3A4 enzyme: involvement of phenylalanine residues in the gating mechanism. *J Phys Chem B* **113,** 13018–13025 (Oct. 2009).
14. Kingsley, L. J. & Lill, M. A. Including ligand-induced protein flexibility into protein tunnel prediction. *J Comput Chem* **35,** 1748–1756 (Sept. 2014).
15. Liu, X., Wang, X. & Jiang, H. A steered molecular dynamics method with direction optimization and its applications on ligand molecule dissociation. *J Biochem Biophys Methods* **70,** 857–64 (Apr. 2008).
16. Ludemann, S. K., Lounnas, V. & Wade, R. C. How do substrates enter and products exit the buried active site of cytochrome P450cam? 2. Steered molecular dynamics and adiabatic mapping of substrate pathways. *J Mol Biol* **303,** 813–830 (Nov. 2000).
17. Urban, P., Truan, G. & Pompon, D. Access channels to the buried active site control substrate specificity in CYP1A P450 enzymes. *Biochimica Et Biophysica Acta-General Subjects* **1850,** 696–707 (Apr. 2015).
18. Rydzewski, J. & Nowak, W. Ligand diffusion in proteins via enhanced sampling in molecular dynamics. *Physics of Life Reviews* **22-23,** 58–74 (Dec. 2017).
19. Ferraro, D., Okerlund, A., Mowers, J. & Ramaswamy, S. Structural basis for regioselectivity and stereoselectivity of product formation by naphthalene 1,2-dioxygenase. *Journal of Bacteriology* **188,** 6986–6994 (Oct. 2006).
20. Ferraro, D. J. *Structure-function Studies of Rieske Oxygenases* ISBN: 054975-2110 (ProQuest, 2008).
21. Ferraro, D. J. *et al.* Structural investigations of the ferredoxin and terminal oxygenase components of the biphenyl 2,3-dioxygenase from Sphingobium yanoikuyae B1. *BMC Struct Biol* **7,** 10 (2007).
22. Ferraro, D. J., Gakhar, L. & Ramaswamy, S. Rieske business: structure–function of Rieske non-heme oxygenases. *Biochemical and biophysical research communications* **338,** 175–190 (2005).
23. Sehnal, D. *et al.* MOLE 2.0: advanced approach for analysis of biomacromolecular channels. *J Cheminform* **5,** 39 (2013).
24. Pravda, L. *et al.* Anatomy of enzyme channels. *BMC Bioinformatics* **15,** 379 (Nov. 2014).
25. Resnick, S., Lee, K. & Gibson, D. Diverse reactions catalyzed by naphthalene dioxygenase from Pseudomonas sp. strain NCIB 9816. *J. Ind. Microbiol.* **17,** 438–457 (Nov. 1996).

26. Carredano, E. *et al.* Substrate binding site of naphthalene 1,2-dioxygenase: functional implications of indole binding. *J Mol Biol* **296,** 701–712 (Feb. 2000).
27. Gibson, D. T. & Parales, R. E. Aromatic hydrocarbon dioxygenases in environmental biotechnology. *Curr Opin Biotechnol* **11,** 236–243 (June 2000).
28. Karlsson, A. *et al.* Crystal structures of naphthalene dioxygenase along the reaction pathway. *Journal of Inorganic Biochemistry* **96,** 164–164 (July 2003).
29. Lee, K., Friemann, R., Parales, J. V., Gibson, D. T. & Ramaswamy, S. Purification, crystallization and preliminary X-ray diffraction studies of the three components of the toluene 2, 3-dioxygenase enzyme system. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications* **61,** 669–672 (2005).
30. Yu, C. L., Parales, R. E. & Gibson, D. T. Multiple mutations at the active site of naphthalene dioxygenase affect regioselectivity and enantioselectivity. *Journal of Industrial Microbiology & Biotechnology* **27,** 94–103 (Aug. 2001).
31. Wackett, L. P. Mechanism and applications of Rieske non-heme iron dioxygenases. *Enzyme and Microbial Technology* **31,** 577–587 (2002).
32. Illanes, A. Enzyme biocatalysis. *Principles and Applications. Editorial Springer-Verlag New York Inc., United States* (2008).
33. Hendrychova, T., Berka, K., Navratilova, V., Anzenbacher, P. & Otyepka, M. Dynamics and hydration of the active sites of mammalian cytochromes P450 probed by molecular dynamics simulations. *Curr Drug Metab* **13,** 177–189 (Feb. 2012).
34. Li, J. *et al.* Electrostatic gating of a nanometer water channel. *Proc Natl Acad Sci U S A* **104,** 3687–92 (Mar. 2007).
35. Yang, K., Liu, X., Wang, X. & Jiang, H. A steered molecular dynamics method with adaptive direction adjustments. *Biochem Biophys Res Commun* **379,** 494–498 (Feb. 2009).
36. Thoden, J. B., Huang, X. Y., Raushel, F. M. & Holden, H. M. Carbamoylphosphate synthetase - Creation of an escape route for ammonia. *J Biol Chem* **277,** 39722–39727 (Oct. 2002).
37. Amaro, R. E., Myers, R. S., Davisson, V. J. & Luthey-Schulten, Z. A. Structural elements in IGP synthase exclude water to optimize ammonia transfer. *Biophysical Journal* **89,** 475–487 (July 2005).
38. Zhou, H. X. & McCammon, J. A. The gates of ion channels and enzymes. *Trends in Biochemical Sciences* **35,** 179–185 (Mar. 2010).
39. Fan, Y. B., Lund, L., Shao, Q., Gao, Y. Q. & Raushel, F. M. A Combined Theoretical and Experimental Study of the Ammonia Tunnel in Carbamoyl Phosphate Synthetase. *J Am Chem Soc* **131,** 10211–10219 (July 2009).

40. Tyzack, J. D., Furnham, N., Sillitoe, I., Orengo, C. M. & Thornton, J. M. Understanding enzyme function evolution from a computational perspective. *Curr Opin Struct Biol* **47,** 131–139 (Dec. 2017).
41. Tyzack, J. D., Williamson, M. J., Torella, R. & Glen, R. C. Prediction of Cytochrome P450 Xenobiotic Metabolism: Tethered Docking and Reactivity Derived from Ligand Molecular Orbital Analysis. *Journal of Chemical Information and Modeling* **53,** 1294–1305 (June 2013).
42. Bassan, A., Blomberg, M. R. A. & Siegbahn, P. E. M. A theoretical study of the cis-dihydroxylation mechanism in naphthalene 1,2-dioxygenase. *Journal of Biological Inorganic Chemistry* **9,** 439–452 (June 2004).
43. Jorgensen, W. L. & Tirado-Rives, J. The Opls Potential Functions for Proteins - Energy Minimizations for Crystals of Cyclic-Peptides and Crambin. *J Am Chem Soc* **110,** 1657–1666 (Mar. 1988).
44. Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J Am Chem Soc* **118,** 11225–11236 (Nov. 1996).
45. McDonald, N. A. & Jorgensen, W. L. Development of an all-atom force field for heterocycles. Properties of liquid pyrrole, furan, diazoles, and oxazoles. *J Phys Chem B* **102,** 8049–8059 (Oct. 1998).
46. Ensley, B. D. *et al.* Expression of naphthalene oxidation genes in Escherichia coli results in the biosynthesis of indigo. *Science* **222,** 167–169 (Oct. 1983).
47. Friemann, R. *et al.* Structures of the multicomponent Rieske non-heme iron toluene 2,3-dioxygenase enzyme system. *Acta Crystallographica Section D Structural Biology* **65,** 24–33 (Jan. 2009).
48. Seo, J., Kang, S. I., Kim, M., Han, J. & Hur, H. G. Flavonoids biotransformation by bacterial non-heme dioxygenases, biphenyl and naphthalene dioxygenase. *Appl Microbiol Biotechnol* **91,** 219–228 (July 2011).
49. Kaushik, S. *et al.* Impact of the access tunnel engineering on catalysis is strictly ligand-specific. *Febs Journal* **285,** 1456–1476 (Apr. 2018).
50. Kauppi, B. *et al.* Structure of an aromatic-ring-hydroxylating dioxygenase naphthalene 1,2-dioxygenase. *Structure* **6,** 571–586 (May 1998).
51. Bassan, A., Blomberg, M. R. A., Siegbahn, P. E. M. & Que, L. Two faces of a biomimetic non-heme HO-Fe-v = O oxidant: Olefin epoxidation versus cis-dihydroxylation. *Angewandte Chemie-International Edition* **44,** 2939–2941 (2005).
52. Costas, M., Mehn, M. P., Jensen, M. P. & Que, J. L. Dioxygen activation at mononuclear nonheme iron active sites: enzymes, models, and intermediates. *Chem Rev* **104,** 939–86 (Feb. 2004).
53. Prokop, Z. *et al. Engineering of protein tunnels: keyhole-lock-key model for catalysis by the enzymes with buried active sites* (Wiley-VCH, Weinheim, 2012).

54. Paloncyova, M., Navratilova, V., Berka, K., Laio, A. & Otyepka, M. Role of Enzyme Flexibility in Ligand Access and Egress to Active Site: Bias-Exchange Metadynamics Study of 1,3,7-Trimethyluric Acid in Cytochrome P450 3A4. *J Chem Theory Comput* **12,** 2101–2109 (Apr. 2016).
55. Liebgott, P. P. *et al.* Relating diffusion along the substrate tunnel and oxygen sensitivity in hydrogenase. *Nature Chemical Biology* **6,** 63–70 (Jan. 2010).
56. Zhou, H. X., Wlodek, S. T. & McCammon, J. A. Conformation gating as a mechanism for enzyme specificity. *Proc Natl Acad Sci U S A* **95,** 9280–9283 (Aug. 1998).
57. Wolfe, M. D., Parales, J. V., Gibson, D. T. & Lipscomb, J. D. Single turnover chemistry and regulation of O2 activation by the oxygenase component of naphthalene 1,2-dioxygenase. *J Biol Chem* **276,** 1945–1953 (Jan. 2001).
58. Olsen, L., Oostenbrink, C. & Jorgensen, F. S. Prediction of cytochrome P450 mediated metabolism. *Adv Drug Deliv Rev* **86,** 61–71 (June 2015).
59. Su, B. H. *et al.* Rule-Based Prediction Models of Cytochrome P450 Inhibition. *J Chem Inf Model* **55,** 1426–1434 (July 2015).
60. Ekins, S., Berbaum, J. & Harrison, R. K. Generation and validation of rapid computational filters for CYP2D6 and CYP3A4. *Drug Metab Dispos* **31,** 1077–1080 (Sept. 2003).
61. Honarparvar, B., Govender, T., Maguire, G. E. M., Soliman, M. E. S. & Kruger, H. G. Integrated Approach to Structure-Based Enzymatic Drug Design: Molecular Modeling, Spectroscopy, and Experimental Bioactivity. *Chem Rev* **114,** 493–537 (Jan. 2014).
62. Sukuru, S. C. K. *et al.* Discovering new classes of Brugia malayi asparaginyl-tRNA synthetase inhibitors and relating specificity to conformational change. *Journal of Computer-Aided Molecular Design* **20,** 159–178 (Mar. 2006).
63. Levitt, D. G. & Banaszak, L. J. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J Mol Graph* **10,** 229–234 (Dec. 1992).
64. Huang, B. & Schroeder, M. LIGSITE csc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct Biol* **6,** 19 (2006).
65. Petrek, M. *et al.* CAVER: a new tool to explore routes from protein clefts, pockets and cavities. *BMC Bioinformatics* **7,** 316 (2006).
66. Yusuf, M. *et al.* H274Y's Effect on Oseltamivir Resistance: What Happens Before the Drug Enters the Binding Site. *J Chem Inf Model* **56,** 82–100 (Jan. 2016).
67. Tran, D. T. T., Le, L. T. & Truong, T. N. Discover binding pathways using the sliding binding-box docking approach: application to binding pathways of oseltamivir to avian influenza H5N1 neuraminidase. *Journal of Computer-Aided Molecular Design* **27,** 689–695 (Aug. 2013).

68. Schramm, V. L. Enzymatic transition states, transition-state analogs, dynamics, thermodynamics, and lifetimes. *Annu Rev Biochem* **80,** 703–732 (2011).
69. Vashisth, H. & Abrams, C. F. Ligand escape pathways and (un)binding free energy calculations for the hexameric insulin-phenol complex. *Biophys J* **95,** 4193–4204 (Nov. 2008).
70. Pavlova, M. *et al.* Redesigning dehalogenase access tunnels as a strategy for degrading an anthropogenic substrate. *Nature Chemical Biology* **5,** 727–733 (Oct. 2009).
71. Grayson, P., Tajkhorshid, E. & Schulten, K. Mechanisms of selectivity in channels and enzymes studied with interactive molecular dynamics. *Biophysical Journal* **85,** 36–48 (July 2003).
72. Skovstrup, S., David, L., Taboureau, O. & Jorgensen, F. S. A Steered Molecular Dynamics Study of Binding and Translocation Processes in the GABA Transporter. *Plos One* **7** (June 2012).
73. Fukunishi, H., Yagi, H., Kamijo, K. & Shimada, J. Role of a Mutated Residue at the Entrance of the Substrate Access Channel in Cytochrome P450 Engineered for Vitamin D-3 Hydroxylation Activity. *Biochemistry* **50,** 8302–8310 (Oct. 2011).
74. Padhi, S. & Priyakumar, U. D. Urea-Aromatic Stacking and Concerted Urea Transport: Conserved Mechanisms in Urea Transporters Revealed by Molecular Dynamics. *Journal of Chemical Theory and Computation* **12,** 5190–5200 (Oct. 2016).
75. Lv, X. Y., Liu, H. H., Ke, M. & Gong, H. P. Exploring the pH-Dependent Substrate Transport Mechanism of FocA Using Molecular Dynamics Simulation. *Biophysical Journal* **105,** 2714–2723 (Dec. 2013).
76. Wang, J. L., Albers, T. & Grewer, C. Energy Landscape of the Substrate Translocation Equilibrium of Plasma-Membrane Glutamate Transporters. *Journal of Physical Chemistry B* **122,** 28–39 (Jan. 2018).
77. Park, M. S. Molecular Dynamics Simulations of the Human Glucose Transporter GLUT1. *Plos One* **10** (Apr. 2015).
78. Irudayanathan, F. J., Wang, N., Wang, X. Y. & Nangia, S. Architecture of the paracellular channels formed by claudins of the blood-brain barrier tight junctions. *Annals of the New York Academy of Sciences* **1405,** 131–146 (Oct. 2017).
79. Luo, Y., Rossi, A. R. & Harris, A. L. Computational Studies of Molecular Permeation through Connexin26 Channels. *Biophysical Journal* **110,** 584–599 (Feb. 2016).
80. Friesner, R. A. *et al.* Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J Med Chem* **49,** 6177–6196 (Oct. 2006).

81. Halgren, T. A. *et al.* Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J Med Chem* **47,** 1750–1759 (Mar. 2004).
82. Friesner, R. A. *et al.* Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem* **47,** 1739–1749 (Mar. 2004).
83. Sastry, G. M., Adzhigirey, M., Day, T., Annabhimoju, R. & Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J Comput Aided Mol Des* **27,** 221–234 (Mar. 2013).
84. Guengerich, F. P. Cytochrome p450 and chemical toxicology. *Chem Res Toxicol* **21,** 70–83 (Jan. 2008).
85. Zuegge, J. *et al.* A fast virtual screening filter for cytochrome P450 3A4 inhibition liability of compound libraries. *Quantitative Structure-Activity Relationships* **21,** 249–256 (Aug. 2002).
86. Larkin, A. *et al.* Application of a fuzzy neural network model in predicting polycyclic aromatic hydrocarbon-mediated perturbations of the CYP1B1 transcriptional regulatory network in mouse skin. *Toxicology and Applied Pharmacology* **267,** 192–199 (Mar. 2013).
87. Molnar, L. & Keseru, G. M. A neural network based virtual screening of cytochrome P450 3A4 inhibitors. *Bioorganic & Medicinal Chemistry Letters* **12,** 419–421 (Feb. 2002).
88. Susnow, R. G. & Dixon, S. L. Use of robust classification techniques for the prediction of human cytochrome P450 2D6 inhibition. *Journal of Chemical Information and Computer Sciences* **43,** 1308–1315 (July 2003).
89. Cheng, F. X. *et al.* Classification of Cytochrome P450 Inhibitors and Noninhibitors Using Combined Classifiers. *J Chem Inf Model* **51,** 996–1011 (May 2011).
90. Yap, C. W. & Chen, Y. Z. Prediction of Cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *Journal of Chemical Information and Modeling* **45,** 982–992 (July 2005).
91. Sun, H. M., Veith, H., Xia, M. H., Austin, C. P. & Huang, R. L. Predictive Models for Cytochrome P450 Isozymes Based on Quantitative High Throughput Screening Data. *J Chem Inf Model* **51,** 2474–2481 (Oct. 2011).
92. Hritz, J., de Ruiter, A. & Oostenbrink, C. Impact of plasticity and flexibility on docking results for cytochrome P450 2D6: a combined approach of molecular dynamics and ligand docking. *J Med Chem* **51,** 7469–77 (Dec. 2008).
93. Teixeira, V. H., Ribeiro, V. & Martel, P. J. Analysis of binding modes of ligands to multiple conformations of CYP3A4. *Biochim Biophys Acta* **1804,** 2036–45 (Oct. 2010).

94. Hayes, C., Ansbro, D. & Kontoyianni, M. Elucidating substrate promiscuity in the human cytochrome 3A4. *J Chem Inf Model* **54,** 857–69 (Mar. 2014).
95. Lee, K., Kauppi, B., Parales, R. E., Gibson, D. T. & Ramaswamy, S. Purification and crystallization of the oxygenase component of naphthalene dioxygenase in native and selenomethionine-derivatized forms. *Biochem Biophys Res Commun* **241,** 553–557 (Dec. 1997).
96. Karlsson, A. *et al.* NO binding to naphthalene dioxygenase. *Journal of Biological Inorganic Chemistry* **10,** 483–489 (Aug. 2005).
97. Wolfe, M. D. & Lipscomb, J. D. Hydrogen peroxide-coupled cis-diol formation catalyzed by naphthalene 1,2-dioxygenase. *J Biol Chem* **278,** 829–835 (Jan. 2003).
98. Chen, Y. C. Beware of docking! *Trends in Pharmacological Sciences* **36,** 78–95 (Feb. 2015).
99. Harder, E. *et al.* OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J Chem Theory Comput* **12,** 281–296 (Jan. 2016).
100. Shivakumar, D. *et al.* Prediction of Absolute Solvation Free Energies using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field. *J Chem Theory Comput* **6,** 1509–1519 (May 2010).
101. Guo, Z. J. *et al.* Probing the alpha-Helical Structural Stability of Stapled p53 Peptides: Molecular Dynamics Simulations and Analysis. *Chemical Biology & Drug Design* **75,** 348–359 (Apr. 2010).
102. Wang, L. L., Berne, B. J. & Friesner, R. A. Ligand binding to protein-binding pockets with wet and dry regions. *Proceedings of the National Academy of Sciences of the United States of America* **108,** 1326–1330 (June 2011).
103. Jorgensen, W. L. & Schyman, P. Treatment of Halogen Bonding in the OPLS-AA Force Field: Application to Potent Anti-HIV Agents. *J Chem Theory Comput* **8,** 3895–3901 (Oct. 2012).
104. Jambeck, J. P. M., Mocci, F., Lyubartsev, A. P. & Laaksonen, A. Partial Atomic Charges and Their Impact on the Free Energy of Solvation. *J Comput Chem* **34,** 187–197 (Jan. 2013).
105. Li, P., Song, L. F. & Merz, J. K. M. Parameterization of highly charged metal ions using the 12-6-4 LJ-type nonbonded model in explicit water. *J Phys Chem B* **119,** 883–895 (Jan. 2015).
106. Lee, J. W., Kim, T. Y., Jang, Y. S., Choi, S. & Lee, S. Y. Systems metabolic engineering for chemicals and materials. *Trends Biotechnol* **29,** 370–378 (Aug. 2011).
107. Park, S. & Schulten, K. Calculating potentials of mean force from steered molecular dynamics simulations. *J Chem Phys* **120,** 5946–5961 (Apr. 2004).

108. Gedeon, P. C., Thomas, J. R. & Madura, J. D. Accelerated Molecular Dynamics and Protein Conformational Change: A Theoretical and Practical Guide Using a Membrane Embedded Model Neurotransmitter Transporter. *Methods in Molecular Biology* **1215,** 253–287 (2015).
109. Furusawa, Y. *et al.* Crystal structure of the terminal oxygenase component of biphenyl dioxygenase derived from Rhodococcus sp. strain RHA1. *J Mol Biol* **342,** 1041–1052 (Sept. 2004).
110. Blomberg, M. R., Borowski, T., Himo, F., Liao, R. Z. & Siegbahn, P. E. Quantum chemical studies of mechanisms for metalloenzymes. *Chem Rev* **114,** 3601–3658 (Apr. 2014).
111. Kovaleva, E. G. & Lipscomb, J. D. Versatility of biological non-heme Fe(II) centers in oxygen activation reactions. *Nat Chem Biol* **4,** 186–193 (Mar. 2008).
112. Peters, M. B. *et al.* Structural Survey of Zinc Containing Proteins and the Development of the Zinc AMBER Force Field (ZAFF). *J Chem Theory Comput* **6,** 2935–2947 (Sept. 2010).
113. Pabis, A., Geronimo, I., York, D. M. & Paneth, P. Molecular Dynamics Simulation of Nitrobenzene Dioxygenase Using AMBER Force Field. *J Chem Theory Comput* **10,** 2246–2254 (June 2014).
114. McGaughey, G. B., Gagne, M. & Rappe, A. K. -stacking interactions - Alive and well in proteins. *J Biol Chem* **273,** 15458–15463 (June 1998).
115. Raunio, H., Kuusisto, M., Juvonen, R. O. & Pentikainen, O. T. Modeling of interactions between xenobiotics and cytochrome P450 (CYP) enzymes. *Frontiers in Pharmacology* **6** (June 2015).
116. Muller, C. S. *et al.* Concurrent Cooperativity and Substrate Inhibition in the Epoxidation of Carbamazepine by Cytochrome P450 3A4 Active Site Mutants Inspired by Molecular Dynamics Simulations. *Biochemistry* **54,** 711–721 (Jan. 2015).
117. Yu, C. L. *et al.* Purification, characterization, and crystallization of the components of a biphenyl dioxygenase system from Sphingobium yanoikuyae B1. *Journal of Industrial Microbiology & Biotechnology* **34,** 311–324 (Apr. 2007).
118. Kumar, P. *et al.* Anaerobic crystallization and initial X-ray diffraction data of biphenyl 2,3-dioxygenase from Burkholderia xenovorans LB400: addition of agarose improved the quality of the crystals. *Acta Crystallographica Section F-Structural Biology Communications* **67,** 59–62 (Jan. 2011).
119. Escalante, D. E., Aukema, K. G., Wackett, L. P. & Aksan, A. Simulation of the Bottleneck Controlling Access into a Rieske Active Site: Predicting Substrates of Naphthalene 1,2-Dioxygenase. *Journal of Chemical Information and Modeling* **57,** 550–561 (Mar. 2017).
120. Frisch, M. J. *et al. Gaussian 09* Gaussian, Inc. (Wallingford, CT, USA, 2009).

121. Kagle, J., Porter, A. W., Murdoch, R. W., Rivera-Cancel, G. & Hay, A. G. Biodegradation of Pharmaceutical and Personal Care Products. *Advances in Applied Mircobiology, Vol 67* **67,** 65–108 (2009).
122. Meynet, P., Head, I. M., Werner, D. & Davenport, R. J. Re-evaluation of dioxygenase gene phylogeny for the development and validation of a quantitative assay for environmental aromatic hydrocarbon degraders. *Fems Microbiology Ecology* **91** (June 2015).
123. Ashikawa, Y. *et al.* Structural insight into the substrate- and dioxygen-binding manner in the catalytic cycle of Rieske nonheme iron oxygenase system, carbazole 1,9a-dioxygenase. *Bmc Structural Biology* **12** (June 2012).
124. Ni Chadhain, S. M., Norman, R. S., Pesce, K. V., Kukor, J. J. & Zystra, G. J. Microbial dioxygenase gene population shifts during polycyclic aromatic hydrocarbon biodegradation. *Applied and Environmental Microbiology* **72,** 4078–4087 (June 2006).
125. Boyd, D. R. *et al.* Dioxygenase-catalysed cis-dihydroxylation of meta-substituted phenols to yield cyclohexenone cis-diol and derived enantiopure cis-triol metabolites. *Organic & Biomolecular Chemistry* **9,** 1479–1490 (2011).
126. Chain, P. S. G. *et al.* Burkholderia xenovorans LB400 harbors a multi-replicon, 9.73-Mbp genome shaped for versatility. *Proceedings of the National Academy of Sciences of the United States of America* **103,** 15280–15287 (Oct. 2006).
127. L'Abbee, J. B., Barriault, D. & Sylvestre, M. Metabolism of dibenzofuran and dibenzo-p-dioxin by the biphenyl dioxygenase of Burkholderia xenovorans LB400 and Comamonas testosteroni B-356. *Applied Microbiology and Biotechnology* **67,** 506–514 (June 2005).
128. Smith, D. J., Park, J., Tiedje, J. M. & Mohn, W. W. A large gene cluster in Burkholderia xenovorans encoding abietane diterpenoid catabolism. *Journal of Bacteriology* **189,** 6195–6204 (Sept. 2007).
129. Gao, J., Ellis, L. B. & Wackett, L. P. The University of Minnesota Pathway Prediction System: multi-level prediction and visualization. *Nucleic Acids Res* **39,** W406–11 (July 2011).
130. Wicker, J., Fenner, K., Ellis, L., Wackett, L. & Kramer, S. Predicting biodegradation products and pathways: a hybrid knowledge- and machine learning-based approach. *Bioinformatics* **26,** 814–821 (Mar. 2010).
131. Escalante, D. E. & Aksan, A. Prediction of Ligand Transport along Hydrophobic Enzyme Nanochannels. *Computational and Structural Biotechnology Journal* (2019).
132. Barry, S. M. & Challis, G. L. Mechanism and Catalytic Diversity of Rieske Non-Heme Iron-Dependent Oxygenases. *Acs Catalysis* **3,** 2362–2370 (Oct. 2013).
133. Copley, S. D. Enzymes with extra talents: moonlighting functions and catalytic promiscuity. *Current Opinion in Chemical Biology* **7,** 265–272 (Apr. 2003).

134. Hermann, J. C. *et al.* Predicting substrates by docking high-energy intermediates to enzyme structures. *J Am Chem Soc* **128,** 15882–91 (Dec. 2006).
135. Koudelakova, T. *et al.* Substrate specificity of haloalkane dehalogenases. *Biochem J* **435,** 345–54 (Apr. 2011).
136. Park, H., Lee, S. & Suh, J. Structural and dynamical basis of broad substrate specificity, catalytic mechanism, and inhibition of cytochrome P450 3A4. *Journal of the American Chemical Society* **127,** 13634–13642 (2005).
137. Schleinkofer, K., Sudarko, Winn, P. J., Ludemann, S. K. & Wade, R. C. Do mammalian cytochrome P450s show multiple ligand access pathways and ligand channelling? *EMBO Rep* **6,** 584–9 (June 2005).
138. Ebert, M. C., Dürr, S. L., A. Houle, A., Lamoureux, G. & Pelletier, J. N. Evolution of P450 monooxygenases toward formation of transient channels and exclusion of nonproductive gases. *Acs Catalysis* **6,** 7426–7437 (2016).
139. Gora, A., Brezovsky, J. & Damborsky, J. Gates of enzymes. *Chem Rev* **113,** 5871–923 (Aug. 2013).
140. Lindberg, R. L. & Negishi, M. Alteration of mouse cytochrome P450coh substrate specificity by mutation of a single amino-acid residue. *Nature* **339,** 632–4 (June 1989).
141. Wang, Y. H., Han, K. L., Yang, S. L. & Yang, L. Structural determinants of steroids for cytochrome P450 3A4-mediated metabolism. *Journal of Molecular Structure-Theochem* **710,** 215–221 (Nov. 2004).
142. Guzik, U., Hupert-Kocurek, K., Sitnik, M. & Wojcieszynska, D. Protocatechuate 3,4-Dioxygenase: A Wide Substrate Specificity Enzyme Isolated from Stenotrophomonas maltophilia KB2 as a Useful Tool in Aromatic Acid Biodegradation. *Journal of Molecular Microbiology and Biotechnology* **24,** 150–160 (2014).
143. Grubmuller, H., Heymann, B. & Tavan, P. Ligand binding: molecular mechanics calculation of the streptavidin-biotin rupture force. *Science* **271,** 997–9 (Feb. 1996).
144. Xu, Y. *et al.* How does huperzine A enter and leave the binding gorge of acetylcholinesterase? Steered molecular dynamics simulations. *Journal of the American Chemical Society* **125,** 11340–11349 (2003).
145. Peplowski, L., Kubiak, K. & Nowak, W. Mechanical aspects of nitrile hydratase enzymatic activity. Steered molecular dynamics simulations of Pseudonocardia thermophila JCM 3095. *Chemical Physics Letters* **467,** 144–149 (2008).
146. Wade, R. C., Winn, P. J. & Schlichting, I. A survey of active site access channels in cytochromes P450. *Journal of inorganic biochemistry* **98,** 1175–1182 (2004).
147. Anandakrishnan, R., Drozdetski, A., Walker, R. C. & Onufriev, A. V. Speed of Conformational Change: Comparing Explicit and Implicit Solvent Molecular Dynamics Simulations. *Biophysical Journal* **108,** 1153–1164 (Mar. 2015).

148. Chakraborty, S., Kumar, H., Dasgupta, C. & Maiti, P. K. Confined Water: Structure, Dynamics, and Thermodynamics. *Acc Chem Res* **50,** 2139–2146 (Sept. 2017).
149. Alexiadis, A. & Kassinos, S. Molecular Simulation of Water in Carbon Nanotubes. *Chemical Reviews* **108,** 5014–5034 (Dec. 2008).
150. Amiri, H., Shepard, K. L., Nuckolls, C. & Hernandez Sanchez, R. Single-Walled Carbon Nanotubes: Mimics of Biological Ion Channels. *Nano Lett* **17,** 1204–1211 (Feb. 2017).
151. Corry, B. Water and ion transport through functionalised carbon nanotubes: implications for desalination technology. *Energy & Environmental Science* **4,** 751–759 (Mar. 2011).
152. Mann, D. J. & Halls, M. D. Water alignment and proton conduction inside carbon nanotubes. *Physical Review Letters* **90** (May 2003).
153. Mao, Z. G. & Sinnott, S. B. A computational study of molecular diffusion and dynamic flow through carbon nanotubes. *Journal of Physical Chemistry B* **104,** 4618–4624 (May 2000).
154. Samoylova, O. N., Calixte, E. I. & Shuford, K. L. Selective ion transport in functionalized carbon nanotubes. *Applied Surface Science* **423,** 154–159 (Nov. 2017).
155. Zhu, F. Q. & Schulten, K. Water and proton conduction through carbon nanotubes as models for biological channels. *Biophysical Journal* **85,** 236–244 (July 2003).
156. Meng, X. W. & Huang, J. P. Enhanced permeation of single-file water molecules across a noncylindrical nanochannel. *Physical Review E* **88** (July 2013).
157. Griffith, J. H. & Scheraga, H. A. Statistical thermodynamics of aqueous solutions - II. Alkali halides at infinite dilution. *Journal of Molecular Structure-Theochem* **711,** 33–48 (Dec. 2004).
158. Wernet, P. *et al.* The structure of the first coordination shell in liquid water. *Science* **304,** 995–999 (May 2004).
159. Czaplewski, C., Liwo, A., Ripoll, D. R. & Scheraga, H. A. Molecular origin of anticooperativity in hydrophobic association. *Journal of Physical Chemistry B* **109,** 8108–8119 (Apr. 2005).
160. Matubayasi, N. & Levy, R. M. Thermodynamics of the hydration shell .2. Excess volume and compressibility of a hydrophobic solute. *Journal of Physical Chemistry* **100,** 2681–2688 (Feb. 1996).
161. Matubayasi, N., Reed, L. H. & Levy, R. M. Thermodynamics of the Hydration Shell .1. Excess Energy of a Hydrophobic Solute. *Journal of Physical Chemistry* **98,** 10640–10649 (Oct. 1994).

162. Setny, P., Baron, R. & McCammon, J. A. How Can Hydrophobic Association Be Enthalpy Driven? *Journal of Chemical Theory and Computation* **6,** 2866–2871 (Sept. 2010).
163. Setny, P. & Geller, M. Water properties inside nanoscopic hydrophobic pocket studied by computer simulations. *Journal of Chemical Physics* **125** (Oct. 2006).
164. Setny, P. *et al.* Dewetting-Controlled Binding of Ligands to Hydrophobic Pockets. *Physical Review Letters* **103** (Oct. 2009).
165. Giovambattista, N., Rossky, P. J. & Debenedetti, P. G. Effect of temperature on the structure and phase behavior of water confined by hydrophobic, hydrophilic, and heterogeneous surfaces. *J Phys Chem B* **113,** 13723–34 (Oct. 2009).
166. Giovambattista, N., Rossky, P. J. & Debenedetti, P. G. Effect of pressure on the phase behavior and structure of water confined between nanoscale hydrophobic and hydrophilic plates. *Physical Review E* **73** (Apr. 2006).
167. Robinson, G. W. *Water in biology, chemistry, and physics : experimental overviews and computational methodologies* 96217259 G. Wilse Robinson ... [et al.]. ill. (some col.) ; 23 cm. Includes bibliographical references (p. 397-463) and indexes. ISBN: 9810224516 (World Scientific, Singapore ; River Edge, NJ, 1996).
168. Bowers, K. J. *et al. Scalable algorithms for molecular dynamics simulations on commodity clusters* in. SC 2006 conference, proceedings of the ACM/IEEE (IEEE, 2006).
169. Reinhard, F. & Grubmuller, H. Estimation of absolute solvent and solvation shell entropies via permutation reduction. *Journal of Chemical Physics* **126** (Jan. 2007).
170. Reinhard, F., Lange, O. F., Hub, J. S., Haas, J. & Grubmuller, H. g_permute: Permutation-reduced phase space density compaction. *Computer Physics Communications* **180,** 455–458 (Mar. 2009).
171. Sasikala, W. D. & Mukherjee, A. Single Water Entropy: Hydrophobic Crossover and Application to Drug Binding. *Journal of Physical Chemistry B* **118,** 10553–10564 (Sept. 2014).
172. Andricioaei, I. & Karplus, M. On the calculation of entropy from covariance matrices of the atomic fluctuations. *Journal of Chemical Physics* **115,** 6289–6292 (Oct. 2001).
173. Lazaridis, T. & Karplus, M. Orientational correlations and entropy in liquid water. *Journal of Chemical Physics* **105,** 4294–4316 (Sept. 1996).
174. Farimani, A. B. & Aluru, N. R. Spatial Diffusion of Water in Carbon Nanotubes: From Fickian to Ballistic Motion. *Journal of Physical Chemistry B* **115,** 12145–12149 (Oct. 2011).
175. Errington, J. R. & Debenedetti, P. G. Relationship between structural order and the anomalies of liquid water. *Nature* **409,** 318–321 (Jan. 2001).

176. Hess, B. Determining the shear viscosity of model liquids from molecular dynamics simulations. *Journal of Chemical Physics* **116,** 209–217 (Jan. 2002).
177. De Marzio, M., Camisasca, G., Conde, M. M., Rovere, M. & Gallo, P. Structural properties and fragile to strong transition in confined water. *J Chem Phys* **146,** 084505 (Feb. 2017).
178. Nieto-Draghi, C., Bonet Avalos, J. & Rousseau, B. Dynamic and structural behavior of different rigid nonpolarizable models of water. *The Journal of chemical physics* **118,** 7954–7964 (2003).
179. Mark, P. & Nilsson, L. Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K. *Journal of Physical Chemistry A* **105,** 9954–9960 (Nov. 2001).
180. Zanotti, J. M. *et al.* Competing coexisting phases in 2D water. *Scientific Reports* **6** (May 2016).
181. Mittal, J., Shen, V. K., Errington, J. R. & Truskett, T. M. Confinement, entropy, and single-particle dynamics of equilibrium hard-sphere mixtures. *Journal of Chemical Physics* **127** (Oct. 2007).
182. Henchman, R. H. Free energy of liquid water from a computer simulation via cell theory. *Journal of Chemical Physics* **126** (Feb. 2007).
183. Witherspoon, P. A. & Saraf, D. N. Diffusion of Methane Ethane Propane and N-Butane in Water from 25 to 43 Degrees. *Journal of Physical Chemistry* **69,** 3752–+ (1965).
184. Phan, A., Cole, D. R., Weiss, R. G., Dzubiella, J. & Striolo, A. Confined Water Determines Transport Properties of Guest Molecules in Narrow Pores. *Acs Nano* **10,** 7646–7656 (Aug. 2016).
185. Kumar, S., Bouzida, D., Swendsen, R. H., Kollman, P. A. & Rosenberg, J. M. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules .1. The Method. *Journal of Computational Chemistry* **13,** 1011–1021 (Oct. 1992).
186. Shaytan, A. K., Shaitan, K. V. & Khokhlov, A. R. Solvent Accessible Surface Area of Amino Acid Residues in Globular Proteins: Correlation of Apparent Transfer Free Energies with Experimental Hydrophobicity Scales. *Biomacromolecules* **10,** 1224–1237 (May 2009).
187. Wheeldon, I. *et al.* Substrate channelling as an approach to cascade reactions. *Nature chemistry* **8,** 299 (2016).
188. Raghunathan, A. V. & Aluru, N. R. Molecular understanding of osmosis in semipermeable membranes. *Phys Rev Lett* **97,** 024501 (July 2006).
189. Brezovsky, J. *et al.* Software tools for identification, visualization and analysis of protein tunnels and channels. *Biotechnol Adv* **31,** 38–49 (Jan. 2013).
190. Macarron, R. *et al.* Impact of high-throughput screening in biomedical research. *Nature Reviews Drug Discovery* **10,** 188–195 (Mar. 2011).

191. Cid, H., Bunster, M., Canales, M. & Gazitua, F. Hydrophobicity and Structural Classes in Proteins. *Protein Engineering* **5,** 373–375 (July 1992).
192. Lonsdale, R., Rouse, S. L., Sansom, M. S. P. & Mulholland, A. J. A Multiscale Approach to Modelling Drug Metabolism by Membrane-Bound Cytochrome P450 Enzymes. *Plos Computational Biology* **10** (July 2014).
193. Das, N. & Chandran, P. Microbial degradation of petroleum hydrocarbon contaminants: an overview. *Biotechnol Res Int* **2011,** 941810 (2011).
194. Atlas, R. M. *et al.* Oil Biodegradation and Oil-Degrading Microbial Populations in Marsh Sediments Impacted by Oil from the Deepwater Horizon Well Blowout. *Environ Sci Technol* **49,** 8356–66 (July 2015).
195. Bailey, L. J. *et al.* Crystallographic analysis of active site contributions to regiospecificity in the diiron enzyme toluene 4-monooxygenase. *Biochemistry* **51,** 1101–13 (Feb. 2012).
196. Dundas, J. *et al.* CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic acids research* **34,** W116–W118 (2006).
197. Lee, K. Benzene-induced uncoupling of naphthalene dioxygenase activity and enzyme inactivation by production of hydrogen peroxide. *Journal of bacteriology* **181,** 2719–2725 (1999).
198. Gore, P. Abnormal Substitution Reactions of Anthracene and Phenanthrene. *The Journal of Organic Chemistry* **22,** 135–138 (1957).
199. Shealy, Y. F. & Clayton, J. D. Synthesis of carbocyclic analogs of purine ribonucleosides. *Journal of the American Chemical Society* **91,** 3075–3083 (1969).
200. Moriya, Y. *et al.* PathPred: an enzyme-catalyzed metabolic pathway prediction server. *Nucleic Acids Res* **38,** W138–43 (July 2010).
201. Ellis, L. B. M., Hershberger, C. D. & Wackett, L. P. The University of Minnesota Biocatalysis/Biodegradation Database: microorganisms, genomics and prediction. *Nucleic Acids Research* **28,** 377–379 (Jan. 2000).
202. Darvas, F. Predicting Metabolic Pathways by Logic Programming. *Journal of Molecular Graphics* **6,** 80–86 (June 1988).
203. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* **38,** D473–9 (Jan. 2010).
204. Jaworska, J., Dimitrov, S., Nikolova, N. & Mekenyan, O. Probabilistic assessment of biodegradability based on metabolic pathways: Catabol system. *Sar and Qsar in Environmental Research* **13,** 307–323 (2002).

Appendices

Appendix A

XML Code

The following is a sample script used for submitting tunnel identification jobs using MOLE 2.0.

```
<?xml version="1.0" encoding="utf-8"?>
<Tunnels>
    <Input SpecificChains="chain">
    directoryTo/inputPDB.pdb</Input>
    <WorkingDirectory>directoryTo/outputResults
    </WorkingDirectory>
    <CustomVdw>
                        <Radius Element="H" Value="1" />
    </CustomVdw>
    <Params ProbeRadius="3.0" InteriorThreshold="1.25"
    SurfaceCoverRadius="10" OriginRadius="5.00"
    BottleneckRadius="0.8" BottleneckLength="3.0"
    BottleneckTolerance="0.0" IgnoreHETAtoms="0" MinDepth="5"/>
    <Export Mesh="1" MeshGz="1" Cavities="1" MeshDensity="0.25"
    PyMol="1" PDB="1" ShortOutput="0" Pores="0"
    PyMolDisplayType="Surface"/>
    <Origin Auto="0">
            <Residue Chain="chain" SequenceNumber="ironNumber"
            />
    </Origin>
</Tunnels>
```

Appendix B

Rotation of molecules

In order to rotate each chemical compound about its center of mass we defined two vectors in the Polar coordinates: $\overrightarrow{V}_{\text{ref}} = (1, 0, 0)$ and $\overrightarrow{V}_{\text{rot}} = (1, 0, 0)$ as shown if Figure B.1. The first vector is used as the reference point from which the rotation will happen. The second vector shows the direction after the rotation transformation.

The cross product of $\overrightarrow{V}_{\text{ref}} = (1, 0, 0)$ and $\overrightarrow{V}_{\text{rot}} = (1, 0, 0)$ defines the vector pointing normal $(\overrightarrow{U})$ to both of them. Therefore we can define a plane orthogonal to vector $(\overrightarrow{U})$ along which the rotation will happen as shown in Figure B.2.

A more detailed explanation of the mathematical operations for the rotation of the molecule, in the context of our algorithm, is described in Appendix C Procedure 3 and §C.2.

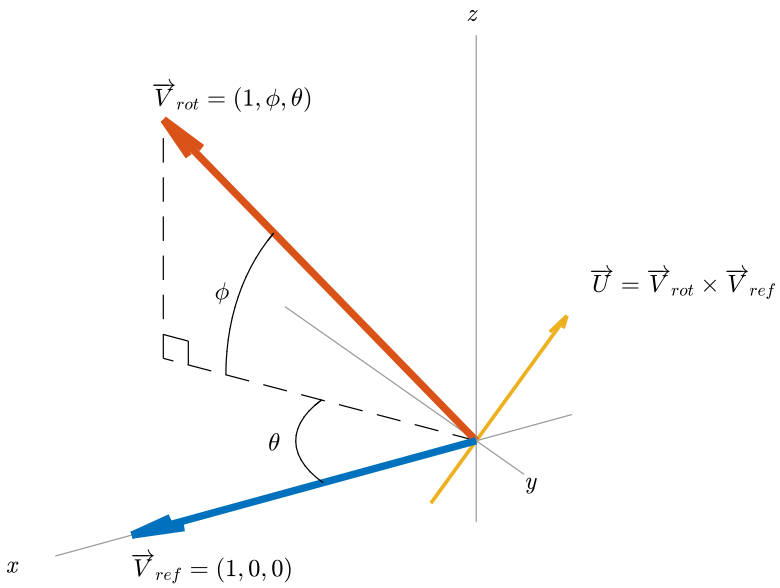

Figure B.1: Definition of the Polar coordinate system used to rotate molecules about angles ϕ and θ.

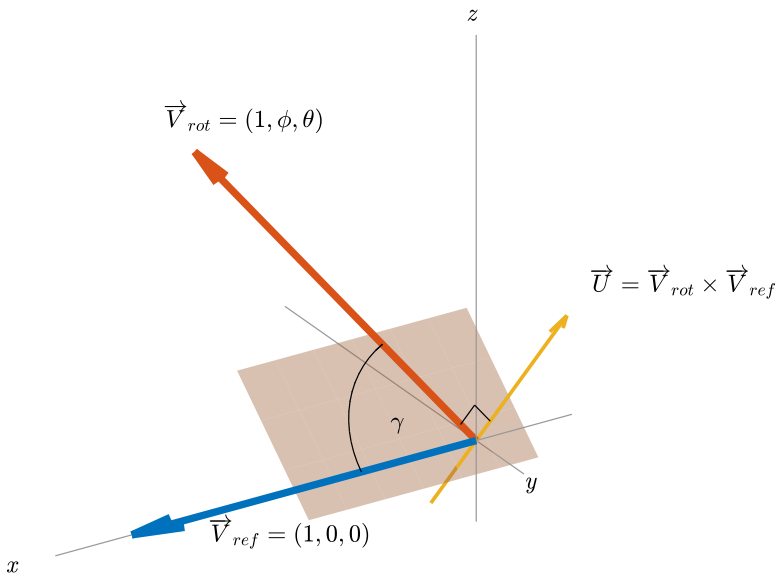

Figure B.2: Plane orthogonal to vector $\overrightarrow{U}$ defines the space about in which the rotation will happen. The angle γ is the distance, in the Polar coordinate system, that the point will be rotated.

Appendix C

Channel Continuity Analysis

C.1 Algorithm Pseudocode

An outline of the developed algorithm is presented in Figure 2.3. Procedures 1-5 provide a more detailed explanation of the method developed.

Blank Space

C.2 Matrix algebra operators

The transformation from Polar coordinates to Cartesian coordinates is described in Procedure 3 lines 2-4. In addition, line 10 of Procedure 3 makes use of the matrix algebra operators described in Equations C.1-C.4.

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{C.1}$$

$$u \otimes u = \begin{bmatrix} u_x^2 & u_x u_y & u_x u_z \\ u_x u_y & u_y^2 & u_y u_z \\ u_x u_z & u_y u_z & u_z^2 \end{bmatrix} \tag{C.2}$$

$$[u]_\times = \begin{bmatrix} 0 & -u_z & u_y \\ u_z & 0 & -u_x \\ -u_y & u_x & 0 \end{bmatrix} \tag{C.3}$$

$$\mathbf{R} = \begin{bmatrix} \cos\gamma + u_x^2(1-\cos\gamma) & u_x u_y(1-\cos\gamma) - u_z \sin\gamma & u_x u_z(1-\cos\gamma) + u_y \sin\gamma \\ u_y u_x(1-\cos\gamma) + u_z \sin\gamma & \cos\gamma + u_y^2(1-\cos\gamma) & u_y u_z(1-\cos\gamma) - u_x \sin\gamma \\ u_z u_x(1-\cos\gamma) - u_y \sin\gamma & u_z u_y(1-\cos\gamma) + u_x \sin\gamma & \cos\gamma + u_z^2(1-\cos\gamma) \end{bmatrix} \tag{C.4}$$

C.3 Calculation of $\Delta G_{\mathbf{A.S}}$ and $\Delta G_{\mathbf{trj}}$

At every step along the channel we calculate the ensemble average free energy of trajectory using Equations 2.3, and 2.5-2.7. Therefore, for every ligand-channel combination, we obtain a free energy of trajectory profile, as shown in Figure 2.7b and Figure C.1(left). The profiles have been fitted with a cubic spline data interpolation using Matlab. The arithmetic average of all these trajectories is calculated in order to obtain a single profile for each ligand-channel combination, shown in Figure C.1(right). The average trajectory free energy for all 45 ligands is shown in Figure C.2.

It is worth pointing out that the scales of both plots has been kept constant to emphasize that although some channels exhibit fairly large free energies (as seen on Figure C.1 left), these are not them norm.

For each of the free energy profiles, we then calculated the average free energy of the trajectory region (ΔG_{trj}) and the average free energy in the active site region ($\Delta G_{\mathrm{A.S}}$), as shown in Figure C.3. The values of r_σ and r_μ are the average docking distance and standard deviation distance, respectively; these two values are obtained from Figure C.5. In Figure C.3, point e is the boundary between the trajectory region and the active site region and is defined as $r_\sigma + r_\mu$. Both values, (ΔG_{trj}) and ($\Delta G_{\mathrm{A.S.}}$), are obtained by integrating under the cubic spline fitting curve. However, for clarity we have only shown the integration area for (ΔG_{trj}), and we simply defined ($\Delta G_{\mathrm{A.S.}}$)

as the average value within the range and highlighted this section of the plot with the thick red line.

C.4 GlideXP docking study

Our attempt to rapidly predict substrates of NDO was focused on using Glide from the Schrödinger suite of software and the GlideXP scoring function. We docked all 45 compound listed in Tables 2.1 and 2.2 in all 100 frames of our MD simulation trajectory sample. The results of the GlideXP docking score are shown in Figure C.4.

C.5 NDO Channel Solvation State

We used the solvate_pocket utility in Desmond to included SPC water molecules inside of the active site cavity of NDO before starting the 40ns molecular dynamic simulation. We used our population sample trajectory frames ($n = 100$) in order to study the spatial distribution of the waters molecules inside the active site. We used Pymol to filter and count all the water molecules that were found to be < 10Å away from the channel centerline at each frame, and Figure 2.9 shows a single snapshot. The statistics of water count and location for all 100 frames is shown in Table 2.7.

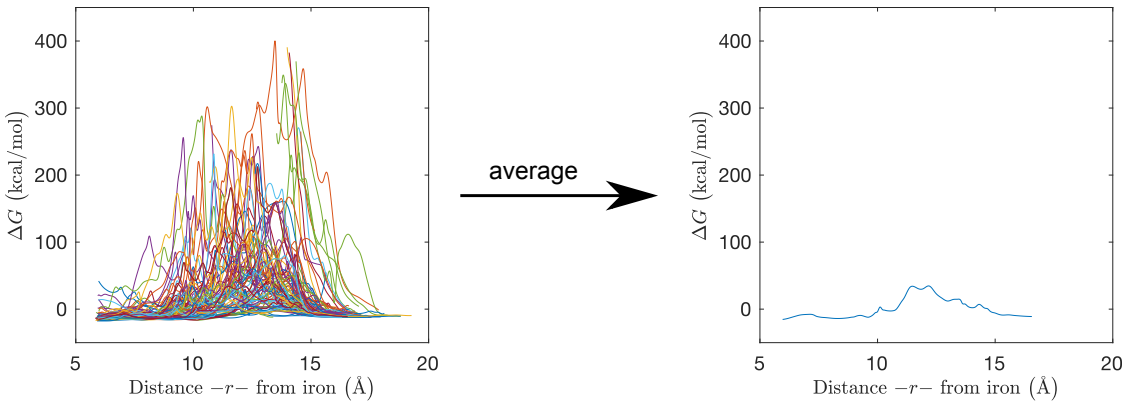

Figure C.1: The left side panel shows the free energy profile along the trajectory of naphthalene in all 100 analyzed frames. The arithmetic average of all frames at each point along the distance –r– is calculated and plotted on the right hand side panel.

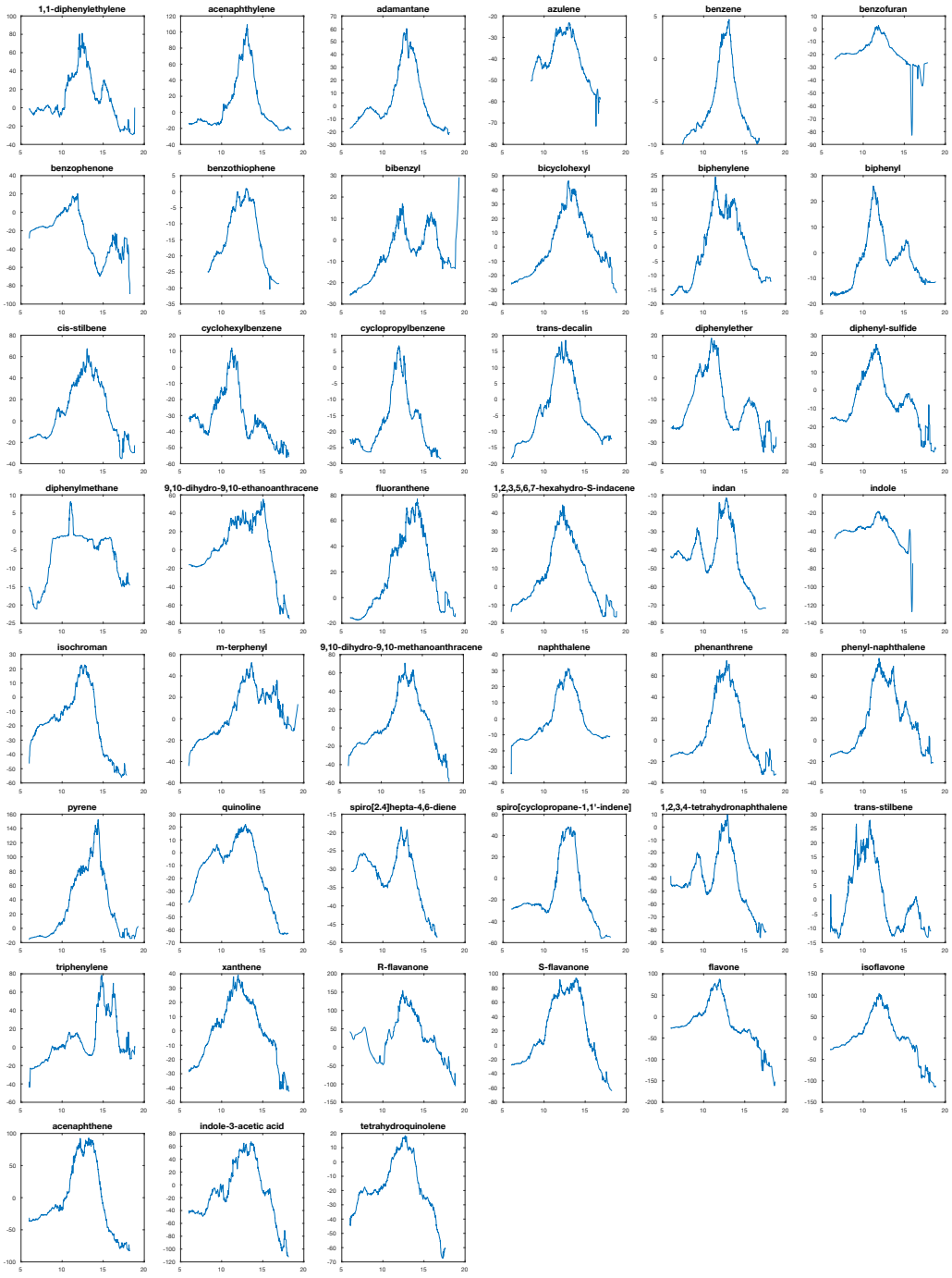

Figure C.2: The left side panel shows the free energy profile along the trajectory of naphthalene in all 100 analyzed frames. The arithmetic average of all frames at each point along the distance –r– is calculated and plotted on the right hand side panel.

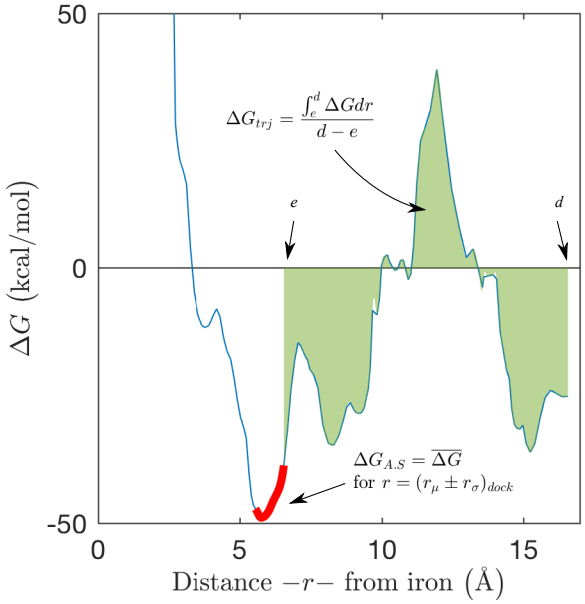

Figure C.3: Full average trajectory free energy profile for biphenyl. The green region shows the integration area under the trajectory profile curve used to calculate ΔG_{trj}. The red line indicates the region of the profile used to calculate $\Delta G_{\text{A.S}}$ The values r_{sigma} and r_{mu} can be obtained from Figure C.5.

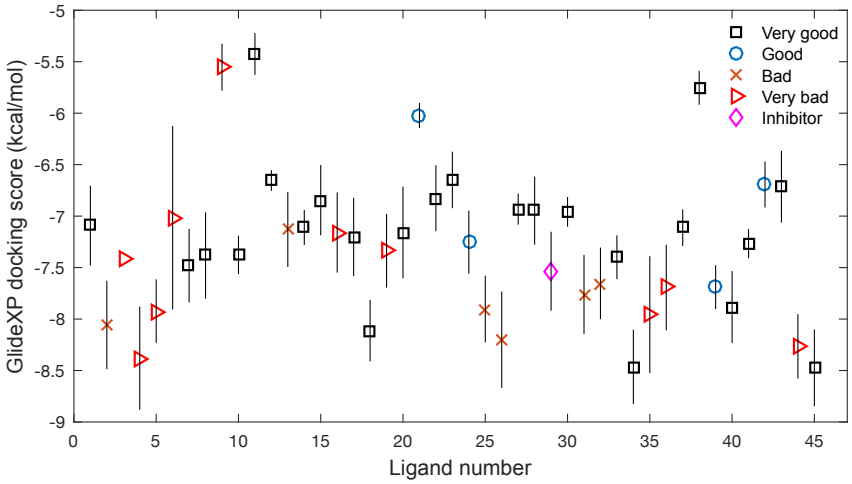

Figure C.4: Docking score distribution using the GlideXP scoring function on all 45 compounds listed in Table S1. The distribution of the docking scores (standard deviation) is shown as the black lines. Compound 3 (9,10-Dihydro-9,10-ethanoanthracene) was only able to dock in 2 (out of 100) grids and both scores resulted in the same value, therefore no standard deviation is reported.

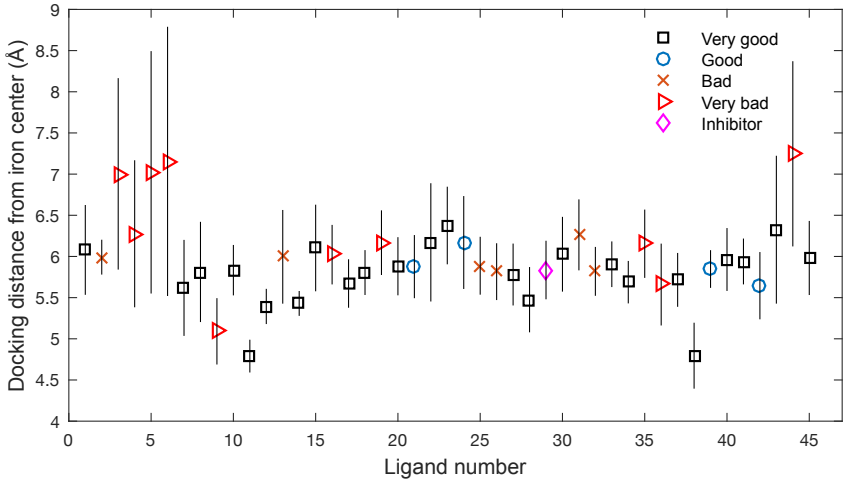

Figure C.5: Docking distance from iron center for all 45 compounds. The distance is defined as the Cartesian distance from the center of mass of the compound to the mononuclear iron center.

Algorithm 1 Preparation and FFP assignment to enzyme trajectory snapshots

```
1: procedure prepareEnzymeAtoms
2:     Prepare crystal structure obtained from PDB.
3:     Build system for Molecular Dynamic Simulation.
4:     Run MD simulation for 50ns.
5:     Extract trajectory snapshots (TS) every 5ps.
6:     Analyze every TS using MOLE 2.0.
7:     if TS contains channel then
8:         Store in TS table.
9:     else
10:        Discard TS.
11:    end if
12:    Select 100 random TS for analysis.
13:    Assign OPLS-AA force field parameters to every atom in all TS.
14:    cL ← Coordinates of channel centerline.
15:    eC ← Coordinates of all atoms in stored TS.
16:    eNB ← Non-bonded parameters for every eC.
17: end procedure
```

Algorithm 2 Preparation and FFP assignment to chemical compound (ligand)

1: **procedure** PREPARECOMPOUNDATOMS
2: Perform energy minimization to optimize geometry.
3: Assign OPLS-AA force field parameters to every atom in the ligand.
4: **cC** ← Coordinates of all the compound atoms
5: **eNB** ← Non-bonded parameters for every **cC**.
6: **end procedure**

Algorithm 3 Rotation Matrix

function ROTATATIONMATRIX(θ, ϕ)
  $x = \cos\theta\cos\phi$
  $y = \cos\theta\sin\phi$
  $z = \sin\theta$
  $\mathbf{V}_{ref} \leftarrow (1, 0, 0)$
  $\mathbf{V}_{rot} \leftarrow (x, y, z)$
  $\mathbf{u} = \mathbf{V}_{ref} \times \mathbf{V}_{rot}$
  $\widehat{\mathbf{u}} = \frac{\mathbf{u}}{\|\mathbf{u}\|}$
  $\gamma = \arccos(\mathbf{V}_{ref} \cdot \mathbf{V}_{rot})$
  $\mathbf{R} = \cos\gamma\mathbf{I} + \sin\gamma[\mathbf{u}]_\times + (1 - cos\gamma)\mathbf{u} \otimes \mathbf{u}$ ▷ See Eqn.C.4
  return R
end function

Algorithm 4 Translation Matrix

```
1: function translationMatrix(from, to)
2:     return (to − from)
3: end function
```

Algorithm 5 Trajectory Analysis

procedure runTrajectoryAnalysis
  cL $\leftarrow$ Coordinates of channel centerline.
  eC $\leftarrow$ Coordinates of all atoms in trajectory snapshots stored.
  eNB $\leftarrow$ Non-bonded parameters for every **eC**.
  cC $\leftarrow$ Coordinates of all the compound atoms.
  eNB $\leftarrow$ Non-bonded parameters for every **cC**.
  $\mathbf{cC_0} =$ translateMatrix(**cC**,$(0,0,0)$)
  $k = 1$ ▷ k is the current path center line step
  for $k \leq k_{stop}$ **do**
    $\mathbf{eC}_k =$translateMatrix($\mathbf{eC}_{k-1}$,$\mathbf{cL}_k$)
    for all θ such that $-\pi/2 < \theta \leq \pi/2$ **do**
      for all ϕ such that $-\pi < \phi \leq \pi$ **do**
        $\mathbf{cC}_R = \mathbf{cC_0} \times$ rotationMatrix(θ, ϕ)
        $i \leftarrow$ atoms in $\mathbf{cC}_R(\theta, \phi)$
        $j \leftarrow$ atoms in $\mathbf{eC}_k(\theta, \phi)$
        $r_{ij} \leftarrow$ distance between atom i and atom j
        $E(\mathbf{x}) = \sum_{i}^{\text{all } i} \sum_{j}^{\text{all } j} \left[\frac{q_i q_j e^2}{r_{ij}} + 4\epsilon_{ij} \left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^{6}}{r_{ij}^{6}} \right) \right]$
        $\rho(\mathbf{x}) = e^{-\beta E(\mathbf{x})}$
      end for
    end for
    $\langle \Delta E_{\text{E-L}} \rangle = \sum_{\theta}^{\text{all } \theta} \sum_{\phi}^{\text{all } \phi} E(\mathbf{x}) \hat{\rho}(\mathbf{x})$ ▷ where $\hat{\rho}$ is the normalized probability
    $S_{\text{config}} = -R \sum_{\theta}^{\text{all } \theta} \sum_{\phi}^{\text{all } \phi} \hat{\rho}(\mathbf{x}) \ln \hat{\rho}(\mathbf{x})$
    $\langle \Delta G \rangle = \langle \Delta E_{\text{E-L}} \rangle - T S_{\text{config}}$
  end for
end procedure